

Total number of questions : 60

13329_DATA ANALYTICS

Time : 1hr

Max Marks : 50

N.B

- 1) All questions are Multiple Choice Questions having single correct option.
- 2) Attempt any 50 questions out of 60.
- 3) Use of calculator is allowed.
- 4) Each question carries 1 Mark.
- 5) Specially abled students are allowed 20 minutes extra for examination.
- 6) Do not use pencils to darken answer.
- 7) Use only black/blue ball point pen to darken the appropriate circle.
- 8) No change will be allowed once the answer is marked on OMR Sheet.
- 9) Rough work shall not be done on OMR sheet or on question paper.
- 10) Darken ONLY ONE CIRCLE for each answer.

Q.no 1. ----- function is used to add a title to each axis instance in a figure.

A : set_title()

B : get_title()

C : set_label()

D : title()

Q.no 2. ----- provides a range of supervised and un-supervised learning algorithms via consistent interface in python

A : Pandas

B : Numpy

C : Scikit-Learn

D : image

Q.no 3. The ----- attribute specifies the number of dimensions or axes of the array.

- A : ndarray.size
- B : ndarray.dtype
- C : ndarray.ndim**
- D : ndarray.axes

Q.no 4. The ----- algorithm is based on the fact that the algorithm uses prior knowledge to find frequent item set.

- A : Clustering
- B : Regression
- C : Naïve Bayes
- D : Apriori**

Q.no 5. ----- submodule of scipy is dedicated to image processing.

- A : ndarray
- B : spatial
- C : ndimage**
- D : special

Q.no 6. If number of input features are 3 then optimal hyperplane in support vector machine is -----

- A : Single point
- B : Line
- C : 2-D Plane**
- D : Non linear line

Q.no 7. ----- is an example of human generated unstructured data.

- A : Text files**
- B : Satellite data
- C : Sensor data

D : Seismic imagery data

Q.no 8. ----- must be installed before you use scikit learn

A : Matlab

B : Scilab

C : Scipy

D : Numpy

Q.no 9. The procedure to organize items of a given collection into groups based on some similar features called as -----

A : Regression

B : Clustering

C : Decision Trees

D : Association

Q.no 10. In statistics, a population consists of -----

A : All People living in a country.

B : All People living in the city.

C : All subjects or objects whose characteristics are being studied.

D : Part of whole dataset

Q.no 11. Which function is used to give title for the axes.

A : plt.title()

B : plt.xlabel()

C : plt.ylabel()

D : plt.xscale()

Q.no 12. ----- function is used to plot a histogram using matplotlib library

A : hist()

B : bar()

C : pie()

D : scatter()

Q.no 13. Which of the following is measure used in decision trees while selecting splitting criteria that partitions data into the best possible manner.

A : Probability

B : Gini Index

C : Regression

D : Association

Q.no 14. Email data is an example of -----

A : Structured data

B : Un-Structured data

C : Semi-Structured data

D : Scattered

Q.no 15. Which of the following is not a type of clustering algorithm?

A : Density clustering

B : K-Mean clustering

C : Centroid clustering

D : Simple clustering

Q.no 16. ----- answers the questions like " How can we make it happen?"

A : Descriptive

B : Prescriptive

C : Predictive

D : Probability

Q.no 17. ----- data does not fits into a data model due to variations in contents.

A : Structured data

B : Un-Structured data

C : Semi-Structured data

D : Scattered

Q.no 18. ----- function multiply two matrices in numpy.

A : prod()

B : mult()

C : dot()

D : *

Q.no 19. ----- is a general purpose array-processing package provides a high performance multi-dimentional array object and tools for working with these arrays.

A : NumPy

B : SciPy

C : sklearn

D : None of these

Q.no 20. ----- library is built on the top of Numpy, SciPy and Matplotlib

A : Sympy

B : Scikit

C : Pandas

D : Numpy

Q.no 21. The last element of ndarray is indexed by -----

A : 0

B : -1

C : 1

D : -2

Q.no 22. -----the step is performed by data scientist after acquiring the data.

A : Data Cleansing

B : Data Integration

C : Data Replication

D : Data loading

Q.no 23. ----- function is used to save an array as in image file.

A : matplotlib.pyplot.image()

B : matplotlib.pyplot.imread()

C : matplotlib.pyplot.imwrite()

D : matplotlib.pyplot.imwrite()

Q.no 24. ----- is unsupervised machine learning technique.

A : KNN

B : Support Vector Machines

C : Decision trees

D : Cluster analysis

Q.no 25. What is correct syntax to generate integers between 10 to 30

A : x=numpy.arange(10,30)

B : x=numpy.array(10,30)

C : x=numpy.arange(10,31)

D : x=arange(10,31)

Q.no 26. ----- function used to get arrays elementwise remainder of division

A : numpy.divide(x1,x2)

B : numpy.mod(x1,x2)

C : numpy.true_divide(x1,x2)

D : numpy.reminder(x1,x2)

Q.no 27. ----- is an indication of how often the rule has been found to be true in association rule mining.

A : Confidence

B : Support

C : Lift

D : None of These

Q.no 28. A ----- is a supervised machine learning algorithm which relies on the assumption of feature independent to classify input data.

A : Clustering

B : Regression

C : Naïve Bayes

D : Apriori

Q.no 29. What is the use of following function? plt.xlabel("Total Marks")

A : Gives label to X-Axis

B : Gives label to Y-Axis

C : Gives title to figure

D : Add text to figure

Q.no 30. Pandas provide ----- function as the entry point for all standard database join operations while merging two DataFrame objects.

A : concat()

B : replace()

C : merge()

D : add()

Q.no 31. Data generated on twitter is an example of -----

A : Structured data

B : Un-Structured data

C : Semi-Structured data

D : Scattered

Q.no 32. ----- is an excellent 2D and 3D graphics library for generating scientific figures?

A : Pandas

B : Numpy

C : matplotlib

D : ndarray

Q.no 33. Support(B) =

A : (Transacions containing (B)) / (Total Transactions)

B : (Transacions containing (B)) / 100

C : (Total Transactions) / (Transacions containing (B))

D : 100/ (Transacions containing (B))

Q.no 34. ----- is an example of semi structured data

A : NoSQL data

B : YouTube data

C : Text File data

D : Satellite imagery data

Q.no 35. ----- is raster graphic format with lossless compression.

A : EPS

B : PDF

C : PNG

D : PS

Q.no 36. -----is a flow-chart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and leaf nodes represent classes or class distributions.

A : Decision tree

B : Association Rule Mining

C : Clustering

D : Support vector machines

Q.no 37. ----- is a form of supervised learning algorithm which is used in mail service providers like Gmail, yahoo, etc. to classify a new mail as spam or

not spam.

A : Classification

B : Regression

C : Clustering

D : Naïve bays

Q.no 38. In ----- the x-axes are grouped into bins and each bin will be treated as a category.

A : Bar

B : Line

C : Scatter

D : Histogram

Q.no 39. When data are collected in a statistical study for only a portion or subset of all elements of interest we are using

A : Sample

B : Parameter

C : Population

D : Probability

Q.no 40. ----- regression finds a relationship between one or more features (independent variables) and a continuous variables (dependent variable).

A : Non-linear

B : Linear

C : Both of these

D : None of These

Q.no 41. It is a measure of disorder or purity or unpredictability or uncertainty.

A : Entropy

B : Support

C : Confidence

D : lift

Q.no 42. Which of the following function is not used to iterate over the rows of the DataFrame.

A : iteritems()

B : iterrows()

C : itertuples()

D : iterpanel()

Q.no 43. ----- is technique that duplicates smaller array to make dimensionality and size of an array as the size and dimensionality of larger array.

A : Multiplation

B : Broadcasting

C : Addition

D : Flatten

Q.no 44. Which of the following task is not performed by Data Scientist.

A : Define the question

B : Create reproducible code

C : Challenge results

D : Staff Recruitement

Q.no 45. To save a figure into a file we can use ----- method in the figure class of matplotlib.pyplot.

A : save()

B : save_fig()

C : Figure()

D : save_image()

Q.no 46. ----- machine learning algorithm used in cross marketing to work with other businesss that complement your own business but not to other competitors.

A : Decision tree

B : Association Rule Mining

C : Clustering

D : Support vector machine

Q.no 47. Which function returns an ndarray object that contains the numbers that are evenly spaced on a log scale.

A : numpy.logspace()

B : numpy.log()

C : numpy.fill()

D : numpy.random()

Q.no 48. The ----- argument of merge function while merging two dataframes specifies which keys are to be included in the resulting dataframe.

A : right

B : on

C : sort

D : how

Q.no 49. Which of the following function is used to split a figure into nrows*ncols sub-axes.

A : plot()

B : draw()

C : bar()

D : subplot()

Q.no 50. ----- function is used to display an image through an external viewer in scipy.

A : display()

B : imread()

C : imshow()

D : show()

Q.no 51. ----- is an unsupervised algorithm used for frequent itemset mining.

A : Apriori

B : Support Vector Machines

C : Decision trees

D : Cluster analysis

Q.no 52. The -- ----- is characterized by a bell shaped curve and area under curve represents probabilities

A : Normal Distribution

B : Binomial Distribution

C : Poission Distribution

D : Probability

Q.no 53. Apriori algorithm uses breadth first search and -----structure to count candidate item sets efficiently.

A : Decision tree

B : Hash tree

C : Red-Black Tree

D : AVL Tree

Q.no 54. In Data science project data acquisition step involves-----

A : Acquiring data from various sources.

B : Selecting dataset

C : Data preprocessing

D : Data modeling

Q.no 55. Select 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 : Analysed data is original source of data.

Q.no 56. Which of the following statement will create an axes at the top right corner of the current figure

A : subplot(2,3,3)

B : subplot(2,3,2)

C : subplot(2,3,4)

D : subplot(2,3,5)

Q.no 57. Catalog design is complex process where the selection of items in a business's catalog are often designed to complement each other so that buying one item will lead to buying of another. So these items are often complements or very related. Which algorithm

A : Decision tree

B : Association Rule Mining

C : Clustering

D : Support vector machine

Q.no 58. While plotting using matplotlib.pyplot A function call similar to subplot(2,3,4) is

A : subplot(234)

B : subplot(243)

C : subplot(324)

D : subplot(4)

Q.no 59. ----- algorithm models a series of logical If-Then- Else decision statements, there is no underlying assumption of a linear or non-linear relationship between the input variables and response variables.

A : Regression

B : Decision Trees

C : Clustering

D : Naïve bays

Q.no 60. To reach to the final point and to make prediction , decision trees must be traversed from -----

A : Top - to - bottom

B : Bottom- to - Top

C : Left- to Right

D : Right - to - Left

Answer for Question No 1. is a

Answer for Question No 2. is c

Answer for Question No 3. is c

Answer for Question No 4. is d

Answer for Question No 5. is c

Answer for Question No 6. is c

Answer for Question No 7. is a

Answer for Question No 8. is c

Answer for Question No 9. is b

Answer for Question No 10. is c

Answer for Question No 11. is a

Answer for Question No 12. is a

Answer for Question No 13. is b

Answer for Question No 14. is b

Answer for Question No 15. is d

Answer for Question No 16. is b

Answer for Question No 17. is b

Answer for Question No 18. is c

Answer for Question No 19. is a

Answer for Question No 20. is b

Answer for Question No 21. is b

Answer for Question No 22. is a

Answer for Question No 23. is d

Answer for Question No 24. is d

Answer for Question No 25. is c

Answer for Question No 26. is b

Answer for Question No 27. is a

Answer for Question No 28. is c

Answer for Question No 29. is a

Answer for Question No 30. is c

Answer for Question No 31. is b

Answer for Question No 32. is c

Answer for Question No 33. is a

Answer for Question No 34. is a

Answer for Question No 35. is c

Answer for Question No 36. is a

Answer for Question No 37. is a

Answer for Question No 38. is d

Answer for Question No 39. is a

Answer for Question No 40. is b

Answer for Question No 41. is a

Answer for Question No 42. is d

Answer for Question No 43. is b

Answer for Question No 44. is d

Answer for Question No 45. is b

Answer for Question No 46. is b

Answer for Question No 47. is a

Answer for Question No 48. is d

Answer for Question No 49. is d

Answer for Question No 50. is c

Answer for Question No 51. is a

Answer for Question No 52. is a

Answer for Question No 53. is b

Answer for Question No 54. is a

Answer for Question No 55. is a

Answer for Question No 56. is a

Answer for Question No 57. is b

Answer for Question No 58. is a

Answer for Question No 59. is b

Answer for Question No 60. is a

Seat No -

Total number of questions : 60

13329_DATA ANALYTICS

Time : 1hr

Max Marks : 50

N.B

- 1) All questions are Multiple Choice Questions having single correct option.
 - 2) Attempt any 50 questions out of 60.
 - 3) Use of calculator is allowed.
 - 4) Each question carries 1 Mark.
 - 5) Specially abled students are allowed 20 minutes extra for examination.
 - 6) Do not use pencils to darken answer.
 - 7) Use only black/blue ball point pen to darken the appropriate circle.
 - 8) No change will be allowed once the answer is marked on OMR Sheet.
 - 9) Rough work shall not be done on OMR sheet or on question paper.
 - 10) Darken ONLY ONE CIRCLE for each answer.
-

Q.no 1. In statistics, a population consists of -----

A : All People living in a country.

B : All People living in the city.

C : All subjects or objects whose characteristics are being studied.

D : Part of whole dataset

Q.no 2. ----- data that depends on data model and resides in a fixed field within a record.

A : Structured data

B : Un-Structured data

C : Semi-Structured data

D : Scattered

Q.no 3. ----- plot displays information as series of data points connected by straight lines.

A : Bar

B : Line

C : Scatter

D : Histogram

Q.no 4. ----- is about developing code to enable the machine to learn to perform tasks and its basic principle is the automatic modeling of underlying that have generated the collected data.

A : Data Science

B : Data Analytics

C : Data Warehousing

D : Data mining

Q.no 5. The ----- function creates a 2-D array with all values 1.

A : **numpy.ones()**

B : **numpy.zeros()**

C : **numpy.eye()**

D : **numpy.empty()**

Q.no 6. ----- method is dataframe reads first n rows from dataframe

A : head(n)

B : tail(n)

C : first(n)

D : start(n)

Q.no 7. Numpy support this function to find trigonometric sine elementwise .

A : **numpy.sin()**

B : **numpy.cosine()**

C : **numpy.tangent()**

D : numpy.rad2sin(x1)

Q.no 8. Apriori algorithm is ----- machine learning algorithm.

A : Un- Supervised

B : Supervised

C : Both of these

D : None of These

Q.no 9. Which library from python is used for implementing machine learning algorithms?

A : Scikit-Learn

B : Pandas

C : Matplotlib

D : Numpy

Q.no 10. The ----- algorithm is based on the fact that the algorithm uses prior knowledge to find frequent item set.

A : Clustering

B : Regression

C : Naïve Bayes

D : Apriori

Q.no 11. Which of the following is not a raster image file format?

A : PNG

B : JPG

C : BMP

D : PDF

Q.no 12. K- nearest neighbors algorithm is based on ----- learning

A : Un- Supervised

B : Supervised

C : Association

D : correlation

Q.no 13. ----- is an example of human generated unstructured data.

A : YouTube data

B : Satellite data

C : Sensor data

D : Seismic imagery data

Q.no 14. Which of the following is NOT supervised learning?

A : PCA

B : Decision Tree

C : Linear Regression

D : Naive Bayesian

Q.no 15. ----- is supervised machine learning algorithm outputs an optimal hyperplane for given labeled training data

A : KNN

B : Support Vector Machines

C : Regression

D : Decision Tree

Q.no 16. ----- rule mining is a technique to identify underlying relations between different items.

A : Classification

B : Regression

C : Clustering

D : Association

Q.no 17. -----type of analytics describes what happened in past

A : Descriptive

B : Prescriptive

C : Predictive

D : Probability

Q.no 18. ----- function is used to add a title to each axis instance in a figure.

A : set_title()

B : get_title()

C : set_label()

D : title()

Q.no 19. Which function is used to give title for the axes.

A : plt.title()

B : plt.xlabel()

C : plt.ylabel()

D : plt.xscale()

Q.no 20. ----- analysis estimates the relationship between single dependent variable and single independent variable

A : Simple Regression

B : Multiple regression

C : Correlation

D : Probability

Q.no 21. In ----- the x-axes are grouped into bins and each bin will be treated as a category.

A : Bar

B : Line

C : Scatter

D : Histogram

Q.no 22. ----- is basic data structure of pandas can be think of SQL table or a spreadsheet data representation.

A : Dataframe

B : series

C : list

D : ndarray

Q.no 23. From matplotlib----- module is used for plotting various plots.

A : Scilearn

B : Pyplot

C : Scilab

D : Matlab

Q.no 24. A perfect negative correlation is signified by -----

A : 1

B : -1

C : 0

D : 2

Q.no 25. ----- is an indication of how often the rule has been found to be true in association rule mining.

A : Confidence

B : Support

C : Lift

D : None of These

Q.no 26. In matplotlib library ----- module supports basic image loading, rescaling and display operations.

A : picture

B : image

C : pyplot

D : sympy

Q.no 27. ----- function from matplotlib.pyplot library plots bar graph for given values of x and y.

A : plot()

B : draw()

C : bar()

D : linedraw()

Q.no 28. ----- is unsupervised technique aiming to divide a multivariate dataset into clusters or groups.

A : KNN

B : Support Vector Machines

C : Regression

D : Cluster analysis

Q.no 29. When data are collected in a statistical study for only a portion or subset of all elements of interest we are using

A : Sample

B : Parameter

C : Population

D : Probability

Q.no 30. ----- is most important language for Data Science.

A : Java

B : Ruby

C : R

D : None of these

Q.no 31. The last element of ndarray is indexed by -----

A : 0

B : -1

C : 1

D : -2

Q.no 32. The number of iterations in apriori -----

A : increases with the size of the data

B : decreases with the increase in size of the data

C : increases with the size of the maximum frequent set

D : decreases with increase in size of the maximum frequent set

Q.no 33. Which of the following is used as attribute selection measure in decision tree algorithms?

A : Information Gain

B : Posterior probability

C : Prior probability

D : Support

Q.no 34. -----is not one of the key data science skill.

A : Statistics

B : Machine Learning

C : Data Visualization

D : software tester

Q.no 35. What is correct syntax to generate integers between 10 to 30

A : x=numpy.arange(10,30)

B : x=numpy.array(10,30)

C : x=numpy.arange(10,31)

D : x=arange(10,31)

Q.no 36. ----- is unsupervised machine learning technique.

A : KNN

B : Support Vector Machines

C : Decision trees

D : Cluster analysis

Q.no 37. ----- searches for the linear optimal separating hyperplane for separation of the data using essential training tuples called support vectors

A : Decision tree

B : Association Rule Mining

C : Clustering

D : Support vector machines

Q.no 38. ----- is a one dimensional array defined in pandas that can be used to store any data type.

A : Dict

B : series

C : ndarray

D : list

Q.no 39. To read image from a file into an array ----- function is used.

A : matplotlib.pyplot.imshow()

B : matplotlib.pyplot.imread()

C : matplotlib.pyplot.imwrite()

D : matplotlib.pyplot.imsave()

Q.no 40. JSON file data is an example of -----

A : Structured data

B : Un-Structured data

C : Semi-Structured data

D : Scattered

Q.no 41. In regression the independent variable is also called as -----

A : Regressor

B : Continuous

C : Regressand

D : Estimated

Q.no 42. ----- function from scipy is used to calculate the distance between all pairs of points in a given set.

A : `scipy.spatial.distance()`

B : `scipy.spatial.distance.measure()`

C : `scipy.spatial.distance.cdist()`

D : `distance(x1,y1)`

Q.no 43. To reach to the final point and to make prediction , decision trees must be traversed from -----

A : Top - to - bottom

B : Bottom- to - Top

C : Left- to Right

D : Right - to - Left

Q.no 44. Which of the following task is not performed by Data Scientist.

A : Define the question

B : Create reproducible code

C : Challenge results

D : Staff Recruitement

Q.no 45. To determine basic salary of a employee when his qualification is given is a ----- problem

A : Correlation

B : Regression

C : Association

D : Qualitative

Q.no 46. Which function from numpy used to return the truncated value of the input elementwise?

A : round()

B : trunc()

C : del()

D : remove_decimal()

Q.no 47. Apriori algorithm uses breadth first search and -----structure to count candidate item sets efficiently.

A : Decision tree

B : Hash tree

C : Red-Black Tree

D : AVL Tree

Q.no 48. While plotting using matplotlib.pyplot A function call similar to subplot(2,3,4) is

A : subplot(234)

B : subplot(243)

C : subplot(324)

D : subplot(4)

Q.no 49. ----- is an unsupervised algorithm used for frequent itemset mining.

A : Apriori

B : Support Vector Machines

C : Decision trees

D : Cluster analysis

Q.no 50. It is a measure of disorder or purity or unpredictability or uncertainty.

A : Entropy

B : Support

C : Confidence

D : lift

Q.no 51. The strength (degree) of the correlation between a set of independent variables X and a dependent variable Y is measured by-----

A : Coefficient of Correlation

B : Coefficient of Determination

C : Standard error of estimate

D : Probability

Q.no 52. To save a figure into a file we can use ----- method in the figure class of matplotlib.pyplot.

A : save()

B : save_fig()

C : Figure()

D : save_image()

Q.no 53. When there is no impact on one variable when increase or decrease on other variable then it is -----

A : Perfect correlation

B : No Correlation

C : Positive Correlation

D : Negative Correlation

Q.no 54. In matplotlib ----- is container class for figure instance.

A : Axes

B : Canvas

C : Figure

D : FigureCanvas

Q.no 55. Plot_number parameter from subplot() function can range from 1 to -----

A : nrows*ncols

B : max

C : nrows

D : ncols

Q.no 56. Which of the following statement will create an axes at the top right corner of the current figure

A : subplot(2,3,3)

B : subplot(2,3,2)

C : subplot(2,3,4)

D : subplot(2,3,5)

Q.no 57. ----- machine learning algorithm used in cross marketing to work with other businesss that complement your own business but not to other competitors.

A : Decision tree

B : Association Rule Mining

C : Clustering

D : Support vector machine

Q.no 58. In unsupervised learning, scikit learn uses ----- method to infer properties of the data.

A : extract()

B : transform()

C : infer()

D : classify()

Q.no 59. In dataframe to compute summary statistics like mean, standard deviation, min and max count etc for each numerical column ----- function is used.

A : display()

B : head()

C : describe()

D : sort()

Q.no 60. The -- ---- is characterized by a bell shaped curve and area under curve represents probabilities

A : Normal Distribution

B : Binomial Distribution

C : Poission Distribution

D : Probability

Answer for Question No 1. is c

Answer for Question No 2. is a

Answer for Question No 3. is b

Answer for Question No 4. is b

Answer for Question No 5. is a

Answer for Question No 6. is a

Answer for Question No 7. is a

Answer for Question No 8. is a

Answer for Question No 9. is a

Answer for Question No 10. is d

Answer for Question No 11. is d

Answer for Question No 12. is b

Answer for Question No 13. is a

Answer for Question No 14. is a

Answer for Question No 15. is b

Answer for Question No 16. is d

Answer for Question No 17. is a

Answer for Question No 18. is a

Answer for Question No 19. is a

Answer for Question No 20. is a

Answer for Question No 21. is d

Answer for Question No 22. is a

Answer for Question No 23. is b

Answer for Question No 24. is c

Answer for Question No 25. is a

Answer for Question No 26. is b

Answer for Question No 27. is c

Answer for Question No 28. is d

Answer for Question No 29. is a

Answer for Question No 30. is c

Answer for Question No 31. is b

Answer for Question No 32. is c

Answer for Question No 33. is a

Answer for Question No 34. is d

Answer for Question No 35. is c

Answer for Question No 36. is d

Answer for Question No 37. is d

Answer for Question No 38. is b

Answer for Question No 39. is b

Answer for Question No 40. is c

Answer for Question No 41. is a

Answer for Question No 42. is c

Answer for Question No 43. is a

Answer for Question No 44. is d

Answer for Question No 45. is b

Answer for Question No 46. is b

Answer for Question No 47. is b

Answer for Question No 48. is a

Answer for Question No 49. is a

Answer for Question No 50. is a

Answer for Question No 51. is a

Answer for Question No 52. is b

Answer for Question No 53. is b

Answer for Question No 54. is d

Answer for Question No 55. is a

Answer for Question No 56. is a

Answer for Question No 57. is b

Answer for Question No 58. is b

Answer for Question No 59. is c

Answer for Question No 60. is a

Seat No -

Total number of questions : 60

13329_DATA ANALYTICS

Time : 1hr

Max Marks : 50

N.B

- 1) All questions are Multiple Choice Questions having single correct option.
 - 2) Attempt any 50 questions out of 60.
 - 3) Use of calculator is allowed.
 - 4) Each question carries 1 Mark.
 - 5) Specially abled students are allowed 20 minutes extra for examination.
 - 6) Do not use pencils to darken answer.
 - 7) Use only black/blue ball point pen to darken the appropriate circle.
 - 8) No change will be allowed once the answer is marked on OMR Sheet.
 - 9) Rough work shall not be done on OMR sheet or on question paper.
 - 10) Darken ONLY ONE CIRCLE for each answer.
-

Q.no 1. ----- analysis estimates the relationship between single dependent variable and single independent variable

A : Simple Regression

B : Multiple regression

C : Correlation

D : Probability

Q.no 2. ----- means part of population chosen for participation in the study

A : Population

B : Sample

C : Association

D : Correlation

Q.no 3. Choose correct option for machine generated unstructured data.

- A : Website data
- B : YouTube data
- C : Text File data
- D : Sensor data**

Q.no 4. To save or write dataframe data into csv file ----- function is used

- A : write_csv()
- B : write_file()
- C : csv_read()
- D : to_csv()**

Q.no 5. ----- uses a tree structure to specify sequences of decisions and consequences.

- A : Regression
- B : Decision trees**
- C : KNN
- D : SVM

Q.no 6. ----- is about developing code to enable the machine to learn to perform tasks and its basic principle is the automatic modeling of underlying that have generated the collected data.

- A : Data Science
- B : Data Analytics**
- C : Data Warehousing
- D : Data mining

Q.no 7. Numpy support this function to find trigonometric sine elementwise .

- A : numpy.sin()**
- B : numpy.cosine()
- C : numpy.tangent()

D : numpy.rad2sin(x1)

Q.no 8. -----type of analytics describes what happened in past

A : Descriptive

B : Prescriptive

C : Predictive

D : Probability

Q.no 9. The ----- algorithm is based on the fact that the algorithm uses prior knowledge to find frequent item set.

A : Clustering

B : Regression

C : Naïve Bayes

D : Apriori

Q.no 10. Satellite image is an example of -----

A : Structured data

B : Un-Structured data

C : Semi-Structured data

D : Scattered

Q.no 11. Unsupervised learning makes sense of ----- data without having any predefined dataset for its training.

A : unlabeled

B : labeled

C : semi-labeled

D : Empty dataset

Q.no 12. Correlation coefficient values lies between---- and --

A : -1 and +1

B : -1 and 0

C : 0 and 1

D : 0 and infinite

Q.no 13. K- nearest neighbors algorithm is based on ----- learning

A : Un- Supervised

B : Supervised

C : Association

D : correlation

Q.no 14. ----- answers the questions like " How can we make it happen?"

A : Descriptive

B : Prescriptive

C : Predictive

D : Probability

Q.no 15. ----- type of plots show all individual data points without connected with lines.

A : Bar

B : Line

C : Scatter

D : Histogram

Q.no 16. ----- chart is a circular plot divides into slices to show numerical proportion.

A : Bar

B : Line

C : Scatter

D : Pie

Q.no 17. Which of the following is measure used in decision trees while selecting splitting criteria that partitions data into the best possible manner.

A : Information Gain

B : Probability

C : Regression

D : Association

Q.no 18. ----- is an example of human generated unstructured data.

A : YouTube data

B : Satellite data

C : Sensor data

D : Seismic imagery data

Q.no 19. ----- charts represents categorical data with rectangular bars

A : Bar

B : Line

C : Scatter

D : Histogram

Q.no 20. In correlation both values are always-----

A : Random

B : sequential

C : Same

D : from same group

Q.no 21. To rotate an image ----- function is used from scipy library.

A : rotation()

B : scipy.move()

C : scipy.ndimage.rotate()

D : scipy.flip()

Q.no 22. A ----- is an example of the most widely used machine learning algorithms much of its popularity is because it can be adapted to almost any type of data.

A : Clustering

B : Regression

C : Decision trees

D : Apriori

Q.no 23. ----- is a classification technique relies on the naïve assumption that input variables are independent of each other.

A : KNN

B : NAïve Bayes

C : Regression

D : Support vector machine

Q.no 24. ----- phase of the data analytics lifecycle usually takes the longest time.

A : Data Preparation

B : Model Planning

C : Model Building

D : Communicate Results

Q.no 25. ----- is an excellent 2D and 3D graphics library for generating scientific figures?

A : Pandas

B : Numpy

C : matplotlib

D : ndarray

Q.no 26. ----- is most important language for Data Science.

A : Java

B : Ruby

C : R

D : None of these

Q.no 27. Which statement will create 5 x 5 array filled with all values 1

A : `x=numpy.ones((5,5))`

B : `x=numpy.ones(5)`

C : `x=numpy.zeros((5,5))`

D : `x=numpy.eye((5,5))`

Q.no 28. Which function returns the identity array with n x n dimension with its main diagonal set to ones and all other elements to zero.

A : `numpy.ones()`

B : `numpy.zeros()`

C : `numpy.fill()`

D : `numpy.identity()`

Q.no 29. From matplotlib----- module is used for plotting various plots.

A : Scilearn

B : Pyplot

C : Scilab

D : Matlab

Q.no 30. In this type of clustering each data type either belongs to a cluster completely or not.

A : Hard clustering

B : Soft Clustering

C : Medium clustering

D : Simple clustering

Q.no 31. ----- function used to add two numpy arrays elementwise.

A : `numpy.add(x1,x2)`

B : `numpy.mod(x1,x2)`

C : `numpy.true_divide(x1,x2)`

D : numpy.addition(x1,x2)

Q.no 32. A -----graph is a circular plot, divided into slices to show numerical proportions.

A : Bar

B : Scatter

C : pie

D : line

Q.no 33. ----- function from matplotlib.pyplot library plots bar graph for given values of x and y.

A : plot()

B : draw()

C : bar()

D : linedraw()

Q.no 34. If a=np.array([1,2,3,4,5,6,7,8,9,10]) then a[2,5,1] will produce output-----

A : 3, 4, 5

B : 3,4,5,6

C : 2,3,4,5

D : 1,2,3,4,5

Q.no 35. Identify the machine generated unstructured data.

A : Website data

B : YouTube data

C : Text File data

D : Satellite imagery data

Q.no 36. -----is not one of the key data science skill.

A : Statistics

B : Machine Learning

C : Data Visualization

D : software tester

Q.no 37. ----- is raster graphic format with lossless compression.

A : EPS

B : PDF

C : PNG

D : PS

Q.no 38. ----- module from sklearn gathers popular unsupervised clustering algorithms.

A : sklearn.covariance

B : sklearn.base

C : sklearn.neighbors

D : sklearn.cluster

Q.no 39. Regression analysis -----

A : Establishes a relationship between two variables

B : Establishes cause and effect

C : Measures growth

D : Measures demand for good

Q.no 40. ----- is an indication of how often the rule has been found to be true in association rule mining.

A : Confidence

B : Support

C : Lift

D : None of These

Q.no 41. The ----- argument of merge function while merging two dataframes specifies which keys are to be included in the resulting dataframe.

A : right

B : on

C : sort

D : how

Q.no 42. Which of the following task is not performed by Data Scientist.

A : Define the question

B : Create reproducible code

C : Challenge results

D : Staff Recruitement

Q.no 43. ----- is an unsupervised algorithm used for frequent itemset mining.

A : Apriori

B : Support Vector Machines

C : Decision trees

D : Cluster analysis

Q.no 44. ----- analysis is a set of statistical processes for estimating the relationships among dependent and independent variables.

A : Regression

B : Decision tree

C : KNN

D : None of These

Q.no 45. While plotting using matplotlib.pyplot A function call similar to subplot(2,3,4) is

A : subplot(234)

B : subplot(243)

C : subplot(324)

D : subplot(4)

Q.no 46. Which of the following statement will create an axes at the top right corner of the current figure

A : subplot(2,3,3)

B : subplot(2,3,2)

C : subplot(2,3,4)

D : subplot(2,3,5)

Q.no 47. ----- function performs the custom operations for the entire dataframe.

A : function()

B : surutine()

C : rutine()

D : pipe()

Q.no 48. It is a measure of disorder or purity or unpredictability or uncertainty.

A : Entropy

B : Support

C : Confidence

D : lift

Q.no 49. Which of the following algorithm is used in Economics, Finance, Biology etc, to model relationships between parameters of interests.

A : Regression

B : Decision Trees

C : Clustering

D : Naïve bays

Q.no 50. The statement subplot(4,3,5) will divide figure into ----- and specify plotting sholud be done on plot number-----

A : 4 x 3, 5

B : 3x 4, 5

C : 3 x 5, 4

D : 5x 3, 4

Q.no 51. The -- ----- is characterized by a bell shaped curve and area under curve represents probabilities

A : Normal Distribution

B : Binomial Distribution

C : Poission Distribution

D : Probability

Q.no 52. ----- is basically extracting particular set of elements from an array.

A : Slicing

B : indexing

C : sorting

D : broadcasting

Q.no 53. In regression the dependent variable is also called as -----

A : Regression

B : Continuous

C : Regressand

D : Independent

Q.no 54. ----- function is used to display an image through an external viewer in scipy.

A : display()

B : imread()

C : imshow()

D : show()

Q.no 55. Plot_number parameter from subplot() function can range from 1 to -----

A : nrows*ncols

B : max

C : nrows

D : ncols

Q.no 56. To reach to the final point and to make prediction , decision trees must be traversed from -----

A : Top - to - bottom

B : Bottom- to - Top

C : Left- to Right

D : Right - to - Left

Q.no 57. Catalog design is complex process where the selection of items in a business's catalog are often designed to complement each other so that buying one item will lead to buying of another. So these items are often complements or very related. Which algorithm

A : Decision tree

B : Association Rule Mining

C : Clustering

D : Support vector machine

Q.no 58. In unsupervised learning, scikit learn uses ----- method to infer properties of the data.

A : extract()

B : transform()

C : infer()

D : classify()

Q.no 59. In dataframe to compute summary statistics like mean, standard deviation, min and max count etc for each numerical column ----- function is used.

A : display()

B : head()

C : describe()

D : sort()

Q.no 60. Which of the following function is used to split a figure into nrow*ncols sub-axes.

- A : plot()
- B : draw()
- C : bar()
- D : subplot()**

Answer for Question No 1. is a

Answer for Question No 2. is b

Answer for Question No 3. is d

Answer for Question No 4. is d

Answer for Question No 5. is b

Answer for Question No 6. is b

Answer for Question No 7. is a

Answer for Question No 8. is a

Answer for Question No 9. is d

Answer for Question No 10. is b

Answer for Question No 11. is a

Answer for Question No 12. is a

Answer for Question No 13. is b

Answer for Question No 14. is b

Answer for Question No 15. is c

Answer for Question No 16. is d

Answer for Question No 17. is a

Answer for Question No 18. is a

Answer for Question No 19. is a

Answer for Question No 20. is a

Answer for Question No 21. is c

Answer for Question No 22. is c

Answer for Question No 23. is b

Answer for Question No 24. is a

Answer for Question No 25. is c

Answer for Question No 26. is c

Answer for Question No 27. is a

Answer for Question No 28. is d

Answer for Question No 29. is b

Answer for Question No 30. is a

Answer for Question No 31. is a

Answer for Question No 32. is c

Answer for Question No 33. is c

Answer for Question No 34. is a

Answer for Question No 35. is d

Answer for Question No 36. is d

Answer for Question No 37. is c

Answer for Question No 38. is d

Answer for Question No 39. is a

Answer for Question No 40. is a

Answer for Question No 41. is d

Answer for Question No 42. is d

Answer for Question No 43. is a

Answer for Question No 44. is a

Answer for Question No 45. is a

Answer for Question No 46. is a

Answer for Question No 47. is d

Answer for Question No 48. is a

Answer for Question No 49. is a

Answer for Question No 50. is a

Answer for Question No 51. is a

Answer for Question No 52. is a

Answer for Question No 53. is c

Answer for Question No 54. is c

Answer for Question No 55. is a

Answer for Question No 56. is a

Answer for Question No 57. is b

Answer for Question No 58. is b

Answer for Question No 59. is c

Answer for Question No 60. is d

Total number of questions : 60

13329_DATA ANALYTICS

Time : 1hr

Max Marks : 50

N.B

- 1) All questions are Multiple Choice Questions having single correct option.
 - 2) Attempt any 50 questions out of 60.
 - 3) Use of calculator is allowed.
 - 4) Each question carries 1 Mark.
 - 5) Specially abled students are allowed 20 minutes extra for examination.
 - 6) Do not use pencils to darken answer.
 - 7) Use only black/blue ball point pen to darken the appropriate circle.
 - 8) No change will be allowed once the answer is marked on OMR Sheet.
 - 9) Rough work shall not be done on OMR sheet or on question paper.
 - 10) Darken ONLY ONE CIRCLE for each answer.
-

Q.no 1. Apriori algorithm is ----- machine learning algorithm.

A : Un- Supervised

B : Supervised

C : Both of these

D : None of These

Q.no 2. CCTV footaage is an example of -----

A : Structured data

B : Un-Structured data

C : Semi-Structured data

D : Scattered

Q.no 3. Choose correct option for machine generated unstructured data.

A : Website data

B : YouTube data

C : Text File data

D : Sensor data

Q.no 4. Pin code of a city is an example of -----

A : Structured data

B : Un-Structured data

C : Semi-Structured data

D : Scattered

Q.no 5. The leaf nodes in decision trees returns the -----

A : decision condition

B : class lables

C : decision on variables

D : test score

Q.no 6. ----- provides arange of supervised and un-supervised learning algorithms via consistant interface in python

A : Pandas

B : Numpy

C : Scikit-Learn

D : image

Q.no 7. To import data from excel file into a dataframe ----- function is provided by pandas package.

A : read_csv()

B : read_file()

C : read()

D : read_excel()

Q.no 8. ----- function used to get positive square root of an numpy array elementwise.

A : `numpy.sqrt(x1)`

B : `numpy.mod(x1)`

C : `numpy.square(x1)`

D : `numpy.find(x1,2)`

Q.no 9. -----function reads an image from a file as an array.

A : `imsave()`

B : `imread()`

C : `read()`

D : None of these

Q.no 10. Numpy support this function to find trigonometric sine elementwise .

A : `numpy.sin()`

B : `numpy.cosine()`

C : `numpy.tangent()`

D : `numpy.rad2sin(x1)`

Q.no 11. In statistics, a population consists of -----

A : All People living in a country.

B : All People living in the city.

C : All subjects or objects whose characteristics are being studied.

D : Part of whole dataset

Q.no 12. In numpy array , array indices always starts from -----

A : 1

B : -1

C : 0

D : 2

Q.no 13. ----- analysis estimates the relationship between single dependent variable and single independent variable

A : Simple Regression

B : Multiple regression

C : Correlation

D : Probability

Q.no 14. ----- refers to the graphical representation of information and data.

A : Data Visualization

B : Data mining

C : Data warehousing

D : Data Structures

Q.no 15. ----- rule mining is a technique to identify underlying relations between different items.

A : Classification

B : Regression

C : Clustering

D : Association

Q.no 16. ----- means part of population chosen for participation in the study

A : Population

B : Sample

C : Association

D : Correlation

Q.no 17. Email data is an example of -----

A : Structured data

B : Un-Structured data

C : Semi-Structured data

D : Scattered

Q.no 18. Probability always lies between ----- and -----

A : 0 and 1

B : -1 and +1

C : -1 and 0

D : 0 and infinite

Q.no 19. Which of the following is not a type of clustering algorithm?

A : Density clustering

B : K-Mean clustering

C : Centroid clustering

D : Simple clustering

Q.no 20. ----- plot displays information as series of data points connected by straight lines.

A : Bar

B : Line

C : Scatter

D : Histogram

Q.no 21. ----- module from sklearn gathers popular unsupervised clustering algorithms.

A : sklearn.covariance

B : sklearn.base

C : sklearn.neighbors

D : sklearn.cluster

Q.no 22. ----- is an example of semi structured data

A : NoSQL data

B : YouTube data

C : Text File data

D : Satellite imagery data

Q.no 23. Which of the following is used as attribute selection measure in decision tree algorithms?

A : Information Gain

B : Posterior probability

C : Prior probability

D : Support

Q.no 24. A -----graph is a circular plot, divided into slices to show numerical proportions.

A : Bar

B : Scatter

C : pie

D : line

Q.no 25. ----- searches for the linear optimal separating hyperplane for separation of the data using essential training tuples called support vectors

A : Decision tree

B : Association Rule Mining

C : Clustering

D : Support vector machines

Q.no 26. -----the step is performed by data scientist after acquiring the data.

A : Data Cleansing

B : Data Integration

C : Data Replication

D : Data loading

Q.no 27. Which function returns the identity array with $n \times n$ dimension with its main diagonal set to ones and all other elements to zero.

A : numpy.ones()

B : numpy.zeros()

C : numpy.fill()

D : numpy.identity()

Q.no 28. ----- function from matplotlib.pyplot library plots bar graph for given values of x and y.

A : plot()

B : draw()

C : bar()

D : linedraw()

Q.no 29. ----- is an excellent 2D and 3D graphics library for generating scientific figures?

A : Pandas

B : Numpy

C : matplotlib

D : ndarray

Q.no 30. The process by which we estimate value of dependent variable on the basis of one or more independent variables is called as -----

A : Correlation

B : Regression

C : Association

D : Qualitative

Q.no 31. A ----- is an example of the most widely used machine learning algorithms much of its popularity is because it can be adapted to almost any type of data.

A : Clustering

B : Regression

C : Decision trees

D : Apriori

Q.no 32. Slop of the regression line of Y on X is also called as

A : Correlation coefficient

B : Regression coefficient

C : Association coefficient

D : Probability

Q.no 33. ----- is the measure of the likeihood that an event will occure in a random experiment

A : Probability

B : Correlation

C : Regression

D : Sample

Q.no 34. What is the use of following function? Plt.xlabel("Total Marks")

A : Gives label to X-Axis

B : Gives label to Y-Axis

C : Gives title to figure

D : Add text to figure

Q.no 35. ----- analysis finds the reasons behind success or failure in past

A : Descriptive

B : Prescriptive

C : Predictive

D : Probability

Q.no 36. Pandas provide ----- function as the entry point for all standard database join operations while merging two DataFrame objects.

A : concat()

B : replace()

C : merge()

D : add()

Q.no 37. JSON file data is an example of -----

A : Structured data

B : Un-Structured data

C : Semi-Structured data

D : Scattered

Q.no 38. Broadcasting is a powerful technique that allows numpy to work with arrays of ----- .

A : Same Shapes

B : Different Shapes

C : Same values

D : Different values

Q.no 39. If scatter diagram is drawn and all scatter points lie on a straight line then it indicates-----

A : No correlation

B : Perfect correlation

C : Regression

D : Skewness

Q.no 40. ----- models search the data space for areas of varied density of data points in the data space.

A : Connectivity models

B : Centroid models

C : Distribution models

D : Density models

Q.no 41. ----- algorithm models a series of logical If-Then- Else decision statements, there is no underlying assumption of a linear or non-linear relationship between the input variables and response variables.

A : Regression

B : Decision Trees

C : Clustering

D : Naïve bays

Q.no 42. In matplotlib ----- is container class for figure instance.

A : Axes

B : Canvas

C : Figure

D : FigureCanvas

Q.no 43. The -- ---- is characterized by a bell shaped curve and area under curve represents probabilities

A : Normal Distribution

B : Binomial Distribution

C : Poission Distribution

D : Probability

Q.no 44. While plotting using matplotlib.pyplot A function call similar to subplot(2,3,4) is

A : subplot(234)

B : subplot(243)

C : subplot(324)

D : subplot(4)

Q.no 45. Catalog design is complex process where the selection of items in a business's catalog are often designed to complement each other so that buying one item will lead to buying of another. So these items are often complements or very related. Which algorithm

A : Decision tree

B : Association Rule Mining

C : Clustering

D : Support vector machine

Q.no 46. To reach to the final point and to make prediction , decision trees must be traversed from -----

A : Top - to - bottom

B : Bottom- to - Top

C : Left- to Right

D : Right - to - Left

Q.no 47. ----- function is used to display an image through an external viewer in scipy.

A : display()

B : imread()

C : imshow()

D : show()

Q.no 48. ----- function performs the custom operations for the entire dataframe.

A : function()

B : surutine()

C : rutine()

D : pipe()

Q.no 49. For testing accuracy of a machine learning algorithm whole data set should be devided into trainin and testing datasets. Which of the following is good proportion for train-test splitting?

A : Train- 70%, Test - 30%

B : Train- 50%, Test - 50%

C : Train- 30%, Test - 70%

D : Train- 100%, Test - 00%

Q.no 50. Which function from numpy used to return the truncated value of the input elementwise?

A : round()

B : trunc()

C : del()

D : remove_decimal()

Q.no 51. When there is no impact on one variable when increase or decrease on other variable then it is -----

A : Perfect correlation

B : No Correlation

C : Positive Correlation

D : Negative Correlation

Q.no 52. Select 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 : Analysed data is original source of data.

Q.no 53. ----- is technique that duplicates smaller array to make dimensionality and size of an array as the size and dimensionality of larger array.

A : Multipliation

B : Broadcasting

C : Addition

D : Flatten

Q.no 54. Apriori algorithm uses breadth first search and -----structure to count candidate item sets efficiently.

A : Decision tree

B : Hash tree

C : Red-Black Tree

D : AVL Tree

Q.no 55. The statement subplot(4,3,5) will divide figure into ----- and specify plotting sholud be done on plot number-----

A : 4 x 3, 5

B : 3x 4, 5

C : 3 x 5, 4

D : 5x 3, 4

Q.no 56. Which of the following task is not performed by Data Scientist.

A : Define the question

B : Create reproducible code

C : Challenge results

D : Staff Recruitement

Q.no 57. Which of the following function is not used to iterate over the rows of the DataFrame.

A : iteritems()

B : iterrows()

C : itertuples()

D : iterpanel()

Q.no 58. Which function returns an ndarray object that contains the numbers that are evenly spaced on a log scale.

A : numpy.logspace()

B : numpy.log()

C : numpy.fill()

D : numpy.random()

Q.no 59. ----- function from scipy is used to calculate the distance between all pairs of points in a given set.

A : scipy.spatial.distance()

B : scipy.spatial.distance.measure()

C : `scipy.spatial.distance.cdist()`

D : `distance(x1,y1)`

Q.no 60. In unsupervised learning, scikit learn uses ----- method to infer properties of the data.

A : `extract()`

B : `transform()`

C : `infer()`

D : `classify()`

Answer for Question No 1. is a

Answer for Question No 2. is b

Answer for Question No 3. is d

Answer for Question No 4. is a

Answer for Question No 5. is b

Answer for Question No 6. is c

Answer for Question No 7. is d

Answer for Question No 8. is a

Answer for Question No 9. is b

Answer for Question No 10. is a

Answer for Question No 11. is c

Answer for Question No 12. is c

Answer for Question No 13. is a

Answer for Question No 14. is a

Answer for Question No 15. is d

Answer for Question No 16. is b

Answer for Question No 17. is b

Answer for Question No 18. is a

Answer for Question No 19. is d

Answer for Question No 20. is b

Answer for Question No 21. is d

Answer for Question No 22. is a

Answer for Question No 23. is a

Answer for Question No 24. is c

Answer for Question No 25. is d

Answer for Question No 26. is a

Answer for Question No 27. is d

Answer for Question No 28. is c

Answer for Question No 29. is c

Answer for Question No 30. is b

Answer for Question No 31. is c

Answer for Question No 32. is b

Answer for Question No 33. is a

Answer for Question No 34. is a

Answer for Question No 35. is a

Answer for Question No 36. is c

Answer for Question No 37. is c

Answer for Question No 38. is b

Answer for Question No 39. is b

Answer for Question No 40. is d

Answer for Question No 41. is b

Answer for Question No 42. is d

Answer for Question No 43. is a

Answer for Question No 44. is a

Answer for Question No 45. is b

Answer for Question No 46. is a

Answer for Question No 47. is c

Answer for Question No 48. is d

Answer for Question No 49. is a

Answer for Question No 50. is b

Answer for Question No 51. is b

Answer for Question No 52. is a

Answer for Question No 53. is b

Answer for Question No 54. is b

Answer for Question No 55. is a

Answer for Question No 56. is d

Answer for Question No 57. is d

Answer for Question No 58. is a

Answer for Question No 59. is c

Answer for Question No 60. is b

Seat No -

Total number of questions : 60

13329_DATA ANALYTICS

Time : 1hr

Max Marks : 50

N.B

- 1) All questions are Multiple Choice Questions having single correct option.
- 2) Attempt any 50 questions out of 60.
- 3) Use of calculator is allowed.
- 4) Each question carries 1 Mark.
- 5) Specially abled students are allowed 20 minutes extra for examination.
- 6) Do not use pencils to darken answer.
- 7) Use only black/blue ball point pen to darken the appropriate circle.
- 8) No change will be allowed once the answer is marked on OMR Sheet.
- 9) Rough work shall not be done on OMR sheet or on question paper.
- 10) Darken ONLY ONE CIRCLE for each answer.

Q.no 1. Email data is an example of -----

A : Structured data

B : Un-Structured data

C : Semi-Structured data

D : Scattered

Q.no 2. The procedure to organize items of a given collection into groups based on some similar features called as -----

A : Regression

B : Clustering

C : Ddecion Trees

D : Association

Q.no 3. ----- is fundamental library used for scientific computing

A : Pandas

B : Numpy

C : Sympy

D : Scipy

Q.no 4. ----- function is used to add a title to each axis instance in a figure.

A : set_title()

B : get_title()

C : set_label()

D : title()

Q.no 5. ----- provides a range of supervised and un-supervised learning algorithms via consistent interface in python

A : Pandas

B : Numpy

C : Scikit-Learn

D : image

Q.no 6. The ----- function creates a 2-D array with diagonal values 1 and rest values zeros.

A : numpy.ones()

B : numpy.zeros()

C : numpy.eye()

D : numpy.empty()

Q.no 7. ----- refers to the graphical representation of information and data.

A : Data Visualization

B : Data mining

C : Data warehousing

D : Data Structures

Q.no 8. To import data from csv file into a dataframe ----- function is provided by pandas package.

A : `read_csv()`

B : `read_file()`

C : `csv_read()`

D : `Frfrom_csv()`

Q.no 9. The ----- function creates a 2-D array with all values 1.

A : `numpy.Ones()`

B : `numpy.zeros()`

C : `numpy.eye()`

D : `numpy.empty()`

Q.no 10. Naïve Bayes is a classification technique based on -----

A : Bayes Theorem

B : Pythagorous Theorum

C : Least square method

D : mean square method

Q.no 11. ----- means part of population chosen for participation in the study

A : Population

B : Sample

C : Association

D : Correlation

Q.no 12. If number of input features are 3 then optimal hyperplane in support vector machine is -----

A : Single point

B : Line

C : 2-D Plane

D : Non linear line

Q.no 13. ----- method is dataframe reads first n rows from dataframe

A : head(n)

B : tail(n)

C : first(n)

D : start(n)

Q.no 14. ----- uses a tree structure to specify sequences of decisions and consequences.

A : Regression

B : Decision trees

C : KNN

D : SVM

Q.no 15. ----- analysis estimates the relationship between single dependent variable and single independent variable

A : Simple Regression

B : Multiple regression

C : Correlation

D : Probability

Q.no 16. ----- library is built on the top of Numpy, SciPy and Matplotlib

A : Sympy

B : Scikit

C : Pandas

D : Numpy

Q.no 17. Which library from python is used for implementing machine learning algorithms?

A : Scikit-Learn

A : Pandas

C : Matplotlib

D : Numpy

Q.no 18. ----- chart is a circular plot divides into slices to show numerical proportion.

A : Bar

B : Line

C : Scatter

D : Pie

Q.no 19. Sattelite image is an example of -----

A : Structured data

B : Un-Structured data

C : Semi-Structured data

D : Scattered

Q.no 20. Which of the following is not a raster image file format?

A : PNG

B : JPG

C : BMP

D : PDF

Q.no 21. Which of the following plots is not used for multidimensional visualization?

A : Andrrews Curves

B : Prallel Chart

C : Deviation Chart

D : Bar

Q.no 22. ----- is the measure of the likeihood that an event will occure in a random experiment

A : Probability

B : Correlation

C : Regression

D : Sample

Q.no 23. The ---- algorithm is the simplest machine learning algorithm, which building the model consists only of storing the training dataset. To make a prediction for a new data point, the algorithm finds the closest data points in the training dataset i.e its

A : Apriori

B : K-Nearest Neighbors

C : K-Means

D : Decision Trees

Q.no 24. If X and Y are both independent of each other, then correlation coefficient is -----

A : 1

B : -1

C : 0

D : 2

Q.no 25. To rotate an image ----- function is used from scipy library.

A : rotation()

B : scipy.move()

C : scipy.ndimage.rotate()

D : scipy.flip()

Q.no 26. To set x Axis lable of a figure----- function is used

A : set_title()

B : set_lable()

C : set_xlabel()

D : get_xlabel()

Q.no 27. In head() / tail() functions of dataframe the default number of elements to display is -----

A : 3

B : 5

C : 1

D : 10

Q.no 28. Regression analysis -----

A : Establishes a relationship between two variables

B : Establishes cause and effect

C : Measures growth

D : Measures demand for good

Q.no 29. ----- is an indication of how frequently the itemset appears in the dataset in association rule mining.

A : Confidence

B : Support

C : Lift

D : None of These

Q.no 30. In decision trees leaf node denotes a -----

A : class distribution

B : test on an attribute

C : outcome of the test

D : class labels

Q.no 31. ----- analysis finds the reasons behind success or failure in past

A : Descriptive

B : Prescriptive

C : Predictive

D : Probability

Q.no 32. In this type of algorithms inputs are provided but not the desired output.

A : Cluster analysis

B : Support Vector Machines

C : Decision trees

D : Naïve bays

Q.no 33. Pandas provide ----- function as the entry point for all standard database join operations while merging two DataFrame objects.

A : concat()

B : replace()

C : merge()

D : add()

Q.no 34. ----- is 2-D data structure defined in pandas in which data arranged in rows and columns.

A : Series

B : Dataframe

C : ndarray

D : list

Q.no 35. ----- is an example of semi structured data

A : NoSQL data

B : YouTube data

C : Text File data

D : Satellite imagery data

Q.no 36. -----the step is performed by data scientist after acquiring the data.

A : Data Cleansing

B : Data Integration

C : Data Replication

D : Data loading

Q.no 37. Entropy is a measure of the randomness in the information being processed.

A : Entropy

B : Support

C : Confidence

D : lift

Q.no 38. The process by which we estimate value of dependent variable on the basis of one or more independent variables is called as -----

A : Correlation

B : Regression

C : Association

D : Qualitative

Q.no 39. ----- is basic data structure of pandas can be think of SQL table or a spreadsheet data representation.

A : Dataframe

B : series

C : list

D : ndarray

Q.no 40. ----- regression finds a relationship between one or more features (independent variables) and a continuous variables (dependent variable).

A : Non-linear

B : Linear

C : Both of these

D : None of These

Q.no 41. Which of the following function is used to split a figure into nrow*ncols sub-axes.

- A : plot()
- B : draw()
- C : bar()
- D : subplot()**

Q.no 42. ----- machine learning algorithm used in cross marketing to work with other businesss that complement your own business but not to other competitors.

- A : Decision tree
- B : Association Rule Mining**
- C : Clustering
- D : Support vector machine

Q.no 43. In datafram to compute summary statistics like mean, standard deviation, min and max count etc for each numerical column ----- function is used.

- A : display()
- B : head()
- C : describe()**
- D : sort()

Q.no 44. Catalog design is complex process where the selection of items in a business's catalog are often designed to complement each other so that buying one item will lead to buying of another. So these items are often complements or very related. Which algorith

- A : Decision tree
- B : Association Rule Mining
- C : Clustering
- D : Support vector machine**

Q.no 45. For testing accuracy of a machine learning algorithm whole data set should be devided into trainin and testing datasets. Which of the following is good preportion for train-test spliting?

A : Train- 70%, Test - 30%

B : Train- 50%, Test - 50%

C : Train- 30%, Test - 70%

D : Train- 100%, Test - 00%

Q.no 46. ----- is basically extracting particular set of elements from an array.

A : Slicing

B : indexing

C : sorting

D : broadcasting

Q.no 47. It is a measure of disorder or purity or unpredictability or uncertainty.

A : Entropy

B : Support

C : Confidence

D : lift

Q.no 48. ----- algorithm models a series of logical If-Then- Else decision statements, there is no underlying assumption of a linear or non-linear relationship between the input variables and response variables.

A : Regression

B : Decision Trees

C : Clustering

D : Naïve bays

Q.no 49. To reach to the final point and to make prediction , decision trees must be traversed from -----

A : Top - to - bottom

B : Bottom- to - Top

C : Left- to Right

D : Right - to - Left

Q.no 50. ----- is an unsupervised algorithm used for frequent itemset mining.

A : Apriori

B : Support Vector Machines

C : Decision trees

D : Cluster analysis

Q.no 51. Which of the following task is not performed by Data Scientist.

A : Define the question

B : Create reproducible code

C : Challenge results

D : Staff Recruitement

Q.no 52. To save a figure into a file we can use ----- method in the figure class of matplotlib.pyplot.

A : save()

B : save_fig()

C : Figure()

D : save_image()

Q.no 53. Plot_number parameter from subplot() function can range from 1 to -----

A : nrows*ncols

B : max

C : nrows

D : ncols

Q.no 54. The -- ---- is characterized by a bell shaped curve and area under curve represents probabilities

A : Normal Distribution

B : Binomial Distribution

C : Poission Distribution

D : Probability

Q.no 55. The statement subplot(4,3,5) will divide figure into ----- and specify plotting sholud be done on plot number-----

A : 4 x 3, 5

B : 3x 4, 5

C : 3 x 5, 4

D : 5x 3, 4

Q.no 56. The strength (degree) of the correlation between a set of independent variables X and a dependent variable Y is measured by-----

A : Coefficient of Correlation

B : Coefficient of Determination

C : Standard error of estimate

D : Probability

Q.no 57. In regression the dependent variable is also called as -----

A : Regression

B : Continuous

C : Regressand

D : Independent

Q.no 58. In matplotlib ----- is container class for figure instance.

A : Axes

B : Canvas

C : Figure

D : FigureCanvas

Q.no 59. Which of the following machine learning algorithm is used for maret basket analysis means to analyze the association of purchased items in a single basket or single purchase.

A : Decision tree

B : Association Rule Mining

C : Clustering

D : Support vector machine

Q.no 60. To determine basic salary of a employee when his qualification is given is a ----- problem

A : Correlation

B : Regression

C : Association

D : Qualitative

Answer for Question No 1. is b

Answer for Question No 2. is b

Answer for Question No 3. is d

Answer for Question No 4. is a

Answer for Question No 5. is c

Answer for Question No 6. is c

Answer for Question No 7. is a

Answer for Question No 8. is a

Answer for Question No 9. is a

Answer for Question No 10. is a

Answer for Question No 11. is b

Answer for Question No 12. is c

Answer for Question No 13. is a

Answer for Question No 14. is b

Answer for Question No 15. is a

Answer for Question No 16. is b

Answer for Question No 17. is a

Answer for Question No 18. is d

Answer for Question No 19. is b

Answer for Question No 20. is d

Answer for Question No 21. is d

Answer for Question No 22. is a

Answer for Question No 23. is b

Answer for Question No 24. is b

Answer for Question No 25. is c

Answer for Question No 26. is c

Answer for Question No 27. is b

Answer for Question No 28. is a

Answer for Question No 29. is b

Answer for Question No 30. is c

Answer for Question No 31. is a

Answer for Question No 32. is a

Answer for Question No 33. is c

Answer for Question No 34. is b

Answer for Question No 35. is a

Answer for Question No 36. is a

Answer for Question No 37. is a

Answer for Question No 38. is b

Answer for Question No 39. is a

Answer for Question No 40. is b

Answer for Question No 41. is d

Answer for Question No 42. is b

Answer for Question No 43. is c

Answer for Question No 44. is b

Answer for Question No 45. is a

Answer for Question No 46. is a

Answer for Question No 47. is a

Answer for Question No 48. is b

Answer for Question No 49. is a

Answer for Question No 50. is a

Answer for Question No 51. is d

Answer for Question No 52. is b

Answer for Question No 53. is a

Answer for Question No 54. is a

Answer for Question No 55. is a

Answer for Question No 56. is a

Answer for Question No 57. is c

Answer for Question No 58. is d

Answer for Question No 59. is b

Answer for Question No 60. is b

Total number of questions : 60

13329_DATA ANALYTICS

Time : 1hr

Max Marks : 50

N.B

- 1) All questions are Multiple Choice Questions having single correct option.
- 2) Attempt any 50 questions out of 60.
- 3) Use of calculator is allowed.
- 4) Each question carries 1 Mark.
- 5) Specially abled students are allowed 20 minutes extra for examination.
- 6) Do not use pencils to darken answer.
- 7) Use only black/blue ball point pen to darken the appropriate circle.
- 8) No change will be allowed once the answer is marked on OMR Sheet.
- 9) Rough work shall not be done on OMR sheet or on question paper.
- 10) Darken ONLY ONE CIRCLE for each answer.

Q.no 1. Numpy support this function to find trigonometric sine elementwise .

A : **numpy.sin()**

B : **numpy.cosine()**

C : **numpy.tangent()**

D : **numpy.rad2sin(x1)**

Q.no 2. SQL record is an example of -----

A : **Structured data**

B : Un-Structured data

C : Semi-Structured data

D : Scattered

Q.no 3. ----- function used to get positive square root of an numpy array elementwise.

A : `numpy.sqrt(x1)`

B : `numpy.mod(x1)`

C : `numpy.square(x1)`

D : `numpy.find(x1,2)`

Q.no 4. ----- data does not fits into a data model due to variatins in contents.

A : Structured data

B : Un-Structured data

C : Semi-Structured data

D : Scattered

Q.no 5. Which of the following is NOT supervised learning?

A : `PCA`

B : Decision Tree

C : Linear Regression

D : Naive Bayesian

Q.no 6. ----- analysis estimates the relationship between single dependent variable and single independent variable

A : Simple Regression

B : Multiple regression

C : Correlation

D : Probability

Q.no 7. Which of the following function is used to create an array of specified shape but filled with random values.

A : `numpy.random.ran()`

B : rank

C : `random.fill()`

D : numpy.fillrandom()

Q.no 8. ----- is an example of human generated unstructured data.

A : YouTube data

B : Satellite data

C : Sensor data

D : Seismic imagery data

Q.no 9. The ----- function creates a 2-D array with all values 1.

A : numpy. Ones()

B : numpy.zeros()

C : numpy.eye()

D : numpy.empty()

Q.no 10. The ----- function creates a 2-D array with all values 0 (zeros).

A : numpy. Ones()

B : numpy.zeros()

C : numpy.eye()

D : numpy.empty()

Q.no 11. ----- is fundamental library used for scientific computing

A : Pandas

B : Numpy

C : Sympy

D : Scipy

Q.no 12. The ----- function creates a 2-D array with diagonal values 1 and rest values zeros.

A : numpy. Ones()

B : numpy.zeros()

C : numpy.eye()

D : numpy.empty()

Q.no 13. Pandas provide ----- method in order to get label based indexing.

A : iloc()

B : loc()

C : ix()

D : xloc()

Q.no 14. The ----- attribute specifies the number of dimensions or axes of the array.

A : ndarray.size

B : ndarray.dtype

C : ndarray.ndim

D : ndarray.axes

Q.no 15. In support vector machines if input features are 2 then the decision boundries or hyperplane is -----.

A : 2-D plane

B : 3-D plane

C : Line

D : point

Q.no 16. -----type of analytics descibes what happened in past

A : Descriptive

B : Prescriptive

C : Predictive

D : Probability

Q.no 17. ---- is an technique to learn from examples and experience, without being explicitly programmed.

A : Machine Learning

B : Software Testing

C : Computer Science

D : Data mining

Q.no 18. ----- means part of population chosen for participation in the study

A : Population

B : Sample

C : Association

D : Correlation

Q.no 19. The ----- algorithm is based on the fact that the algorithm uses prior knowledge to find frequent item set.

A : Clustering

B : Regression

C : Naïve Bayes

D : Apriori

Q.no 20. ----- chart is a circular plot divided into slices to show numerical proportion.

A : Bar

B : Line

C : Scatter

D : Pie

Q.no 21. -----is a flow-chart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and leaf nodes represent classes or class distributions.

A : Decision tree

B : Association Rule Mining

C : Clustering

D : Support vector machines

Q.no 22. What is correct syntax to generate integers between 10 to 30

A : x=numpy.arange(10,30)

B : x=numpy.array(10,30)

C : x=numpy.arange(10,31)

D : x=arange(10,31)

Q.no 23. ----- is an indication of how often the rule has been found to be true in association rule mining.

A : Confidence

B : Support

C : Lift

D : None of These

Q.no 24. ----- function is used to save an array as in image file.

A : matplotlib.pyplot.image()

B : matplotlib.pyplot.imread()

C : matplotlib.pyplot.imwrite()

D : matplotlib.pyplot.imsave()

Q.no 25. If X and Y are both independent of each other, then correlation coefficient is -----

A : 1

B : -1

C : 0

D : 2

Q.no 26. JSON file data is an example of -----

A : Structured data

B : Un-Structured data

C : Semi-Structured data

D : Scattered

Q.no 27. What is the use of following function? Plt.xlabel("Total Marks")

A : Gives label to X-Axis

B : Gives label to Y-Axis

C : Gives title to figure

D : Add text to figure

Q.no 28. Regression analysis -----

A : Establishes a relationship between two variables

B : Establishes cause and effect

C : Measures growth

D : Measures demand for good

Q.no 29. In this type of algorithms inputs are provided but not the desired output.

A : Cluster analysis

B : Support Vector Machines

C : Decision trees

D : Naïve bays

Q.no 30. ----- analysis finds the reasons behind success or failure in past

A : Descriptive

B : Prescriptive

C : Predictive

D : Probability

Q.no 31. ----- models search the data space for areas of varied density of data points in the data space.

A : Connectivity models

B : Centroid models

C : Distribution models

D : Density models

Q.no 32. ----- function used to get arrays elementwise remainder of division

A : numpy.divide(x1,x2)

B : numpy.mod(x1,x2)

C : numpy.true_divide(x1,x2)

D : numpy.reminder(x1,x2)

Q.no 33. If a=np.array([1,2,3,4,5,6,7,8,9,10]) then a[2,5,1] will produce output-----

A : 3, 4, 5

B : 3,4,5,6

C : 2,3,4,5

D : 1,2,3,4,5

Q.no 34. Slop of the regression line of Y on X is also called as

A : Correlation coefficient

B : Regression coefficient

C : Association coefficient

D : Probability

Q.no 35. The process by which we estimate value of dependent variable on the basis of one or more independent variables is called as -----

A : Correlation

B : Regression

C : Association

D : Qualitative

Q.no 36. In head() / tail() functions of dataframe the default number of elements to display is -----

A : 3

B : 5

C : 1

D : 10

Q.no 37. A perfect negative correlation is signified by -----

A : 1

B : -1

C : 0

D : 2

Q.no 38. ----- is unsupervised technique aiming to divide a multivariate dataset into clusters or groups.

A : KNN

B : Support Vector Machines

C : Regression

D : Cluster analysis

Q.no 39. Among the following clustering algorithm types in which of the following type the notion of similarity is derived by the closeness of a data point to the centroid of the clusters.

A : Connectivity models

B : Centroid models

C : Distribution models

D : Density models

Q.no 40. ----- is an example of semi structured data

A : XML data

B : YouTube data

C : Text File data

D : Satellite imagery data

Q.no 41. Plot_number parameter from subplot() function can range from 1 to -----

A : nrows*ncols

B : max

C : nrows

D : ncols

Q.no 42. The -- ---- is characterized by a bell shaped curve and area under curve represents probabilities

A : Normal Distribution

B : Binomial Distribution

C : Poission Distribution

D : Probability

Q.no 43. Which of the following function is used to split a figure into nrows*ncols sub-axes.

A : plot()

B : draw()

C : bar()

D : subplot()

Q.no 44. ----- is an unsupervised algorithm used for frequent itemset mining.

A : Apriori

B : Support Vector Machines

C : Decision trees

D : Cluster analysis

Q.no 45. ----- analysis is a set of statistical processes for estimating the relationships among dependent and independent variables.

A : Regression

B : Decision tree

C : KNN

D : None of These

Q.no 46. To determine basic salary of a employee when his qualification is given is a ----- problem

A : Correlation

B : Regression

C : Association

D : Qualitative

Q.no 47. In Data science project data acquisition step involves-----

A : Acquiring data from various sources.

B : Selecting dataset

C : Data preprocessing

D : Data modeling

Q.no 48. ----- is technique that duplicates smaller array to make dimensionality and size of an array as the size and dimensionality of larger array.

A : Multiplation

B : Broadcasting

C : Addition

D : Flatten

Q.no 49. Which function from numpy used to return the truncated value of the input elementwise?

A : round()

B : trunc()

C : del()

D : remove_decimal()

Q.no 50. ----- function is used to display an image through an external viewer in scipy.

A : display()

B : imread()

C : imshow()

D : show()

Q.no 51. Which of the following machine learning algorithm is used for market basket analysis means to analyze the association of purchased items in a single basket or single purchase.

A : Decision tree

B : Association Rule Mining

C : Clustering

D : Support vector machine

Q.no 52. ----- machine learning algorithm used in cross marketing to work with other businesss that complement your own business but not to other competitors.

A : Decision tree

B : Association Rule Mining

C : Clustering

D : Support vector machine

Q.no 53. In regression the independent variable is also called as -----

A : Regressor

B : Continuous

C : Regressand

D : Estimated

Q.no 54. ----- algorithm models a series of logical If-Then- Else decision statements, there is no underlying assumption of a linear or non-linear relationship between the input variables and response variables.

A : Regression

B : Decision Trees

C : Clustering

D : Naïve bays

Q.no 55. It is a measure of disorder or purity or unpredictability or uncertainty.

A : Entropy

B : Support

C : Confidence

D : lift

Q.no 56. Which of the following statement will create an axes at the top right corner of the current figure

A : subplot(2,3,3)

B : subplot(2,3,2)

C : subplot(2,3,4)

D : subplot(2,3,5)

Q.no 57. The ----- argument of merge function while merging two dataframes specifies which keys are to be included in the resulting dataframe.

A : right

B : on

C : sort

D : how

Q.no 58. In regression the dependent variable is also called as -----

A : Regression

B : Continuous

C : Regressand

D : Independent

Q.no 59. To save a figure into a file we can use ----- method in the figure class of matplotlib.pyplot.

A : save()

B : save_fig()

C : Figure()

D : save_image()

Q.no 60. Which of the following function is not used to iterate over the rows of the DataFrame.

A : iteritems()

B : iterrows()

C : itertuples()

D : iterpanel()

Answer for Question No 1. is a

Answer for Question No 2. is a

Answer for Question No 3. is a

Answer for Question No 4. is b

Answer for Question No 5. is a

Answer for Question No 6. is a

Answer for Question No 7. is a

Answer for Question No 8. is a

Answer for Question No 9. is a

Answer for Question No 10. is b

Answer for Question No 11. is d

Answer for Question No 12. is c

Answer for Question No 13. is b

Answer for Question No 14. is c

Answer for Question No 15. is c

Answer for Question No 16. is a

Answer for Question No 17. is a

Answer for Question No 18. is b

Answer for Question No 19. is d

Answer for Question No 20. is d

Answer for Question No 21. is a

Answer for Question No 22. is c

Answer for Question No 23. is a

Answer for Question No 24. is d

Answer for Question No 25. is b

Answer for Question No 26. is c

Answer for Question No 27. is a

Answer for Question No 28. is a

Answer for Question No 29. is a

Answer for Question No 30. is a

Answer for Question No 31. is d

Answer for Question No 32. is b

Answer for Question No 33. is a

Answer for Question No 34. is b

Answer for Question No 35. is b

Answer for Question No 36. is b

Answer for Question No 37. is c

Answer for Question No 38. is d

Answer for Question No 39. is b

Answer for Question No 40. is a

Answer for Question No 41. is a

Answer for Question No 42. is a

Answer for Question No 43. is d

Answer for Question No 44. is a

Answer for Question No 45. is a

Answer for Question No 46. is b

Answer for Question No 47. is a

Answer for Question No 48. is b

Answer for Question No 49. is b

Answer for Question No 50. is c

Answer for Question No 51. is b

Answer for Question No 52. is b

Answer for Question No 53. is a

Answer for Question No 54. is b

Answer for Question No 55. is a

Answer for Question No 56. is a

Answer for Question No 57. is d

Answer for Question No 58. is c

Answer for Question No 59. is b

Answer for Question No 60. is d

Seat No -

Total number of questions : 60

13329_DATA ANALYTICS

Time : 1hr

Max Marks : 50

N.B

- 1) All questions are Multiple Choice Questions having single correct option.
- 2) Attempt any 50 questions out of 60.
- 3) Use of calculator is allowed.
- 4) Each question carries 1 Mark.
- 5) Specially abled students are allowed 20 minutes extra for examination.
- 6) Do not use pencils to darken answer.
- 7) Use only black/blue ball point pen to darken the appropriate circle.
- 8) No change will be allowed once the answer is marked on OMR Sheet.
- 9) Rough work shall not be done on OMR sheet or on question paper.
- 10) Darken ONLY ONE CIRCLE for each answer.

Q.no 1. Unsupervised learning makes sense of ----- data without having any predefined dataset for its training.

A : unlabeled

B : labeled

C : semi-labeled

D : Empty dataset

Q.no 2. For multidimensional visualization ----- are used.

A : pie charts

B : Bar charts

C : Andrews curves

D : Scatter plots

Q.no 3. ----- refers to the graphical representation of information and data.

A : Data Visualization

B : Data mining

C : Data warehousing

D : Data Structures

Q.no 4. ----- function multiply two matrices in numpy.

A : prod()

B : mult()

C : dot()

D : *

Q.no 5. If number of input features are 3 then optimal hyperplane in support vector machine is -----

A : Single point

B : Line

C : 2-D Plane

D : Non linear line

Q.no 6. Probability always lies between ---- and ----

A : 0 and 1

B : -1 and +1

C : -1 and 0

D : 0 and infinite

Q.no 7. ----- answers the questions like " How can we make it happen?"

A : Descriptive

B : Prescriptive

C : Predictive

D : Probability

Q.no 8. Pandas provide ----- method in order to get label based indexing.

A : iloc()

B : loc()

C : ix()

D : xloc()

Q.no 9. ----- analysis estimates the relationship between single dependent variable and single independent variable

A : Simple Regression

B : Multiple regression

C : Correlation

D : Probability

Q.no 10. ----- is a general purpose array-processing package provides a high performance multi-dimentional array object and tools for working with these arrays.

A : NumPy

B : SciPy

C : sklearn

D : None of these

Q.no 11. The leaf nodes in decision trees returns the -----

A : decision condition

B : class lables

C : decision on variables

D : test score

Q.no 12. Sattelite image is an example of -----

A : Structured data

B : Un-Structured data

C : Semi-Structured data

D : Scattered

Q.no 13. The ----- function creates a 2-D array with all values 0 (zeros).

A : numpy.ones()

B : numpy.zeros()

C : numpy.eye()

D : numpy.empty()

Q.no 14. ----- function used to get positive square root of an numpy array elementwise.

A : numpy.sqrt(x1)

B : numpy.mod(x1)

C : numpy.square(x1)

D : numpy.find(x1,2)

Q.no 15. Pin code of a city is an example of -----

A : Structured data

B : Un-Structured data

C : Semi-Structured data

D : Scattered

Q.no 16. ----- is fundamental library used for scientific computing

A : Pandas

B : Numpy

C : Sympy

D : Scipy

Q.no 17. Find odd one out from the following :

A : KNN

B : NAïve Bayes

C : Decision Trees

D : Cluster analysis

Q.no 18. ----- is supervised machine learning algorithm outputs an optimal hyperplane for given labeled training data

A : KNN

B : Support Vector Machines

C : Regression

D : Decision Tree

Q.no 19. To import data from csv file into a dataframe ----- function is provided by pandas package.

A : read_csv()

B : read_file()

C : csv_read()

D : Frrom_csv()

Q.no 20. SQL record is an example of -----

A : Structured data

B : Un-Structured data

C : Semi-Structured data

D : Scattered

Q.no 21. JSON file data is an example of -----

A : Structured data

B : Un-Structured data

C : Semi-Structured data

D : Scattered

Q.no 22. ----- is most important language for Data Science.

A : Java

B : Ruby

C : R

D : None of these

Q.no 23. ----- is 2-D data structure defined in pandas in which data arranged in rows and columns.

A : Series

B : Dataframe

C : ndarray

D : list

Q.no 24. -----is a flow-chart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and leaf nodes represent classes or class distributions.

A : Decision tree

B : Association Rule Mining

C : Clustering

D : Support vector machines

Q.no 25. Which of the following is not used for 2-D Visualisation?

A : pie charts

B : Bar charts

C : Andrews curves

D : Scatter plots

Q.no 26. The ----- of a numpy array is a tuple of integers giving the size of the array along each dimension.

A : axes

B : rank

C : shape

D : size

Q.no 27. Pandas provide ----- method in order to get purly integer based indexing.

A : iloc()

B : loc()

C : ix()

D : xloc()

Q.no 28. ----- in decision tree measures how much information a feature gives us about the class

A : Information Gain

B : Posterior probability

C : Prior probability

D : probability

Q.no 29. The process by which we estimate value of dependent variable on the basis of one or more independent variables is called as -----

A : Correlation

B : Regression

C : Association

D : Qualitative

Q.no 30. ----- module from sklearn gathers popular unsupervised clustering algorithms.

A : sklearn.covariance

B : sklearn.base

C : sklearn.neighbors

D : sklearn.cluster

Q.no 31. A ----- is a supervised machine learning algorithm which relies on the assumption of feature independent to classify input data.

A : Clustering

B : Regression

C : Naïve Bayes

D : Apriori

Q.no 32. ----- is a form of supervised learning algorithm which is used in mail service providers like Gmail, yahoo, etc. to classify a new mail as spam or not spam.

A : Classification

B : Regression

C : Clustering

D : Naïve bays

Q.no 33. The objective of ----- algorithm is to find a hyperplane in an N-dimensional space that distinctly classifies the data points.

A : KNN

B : Support Vector Machines

C : Regression

D : Decision Tree

Q.no 34. ----- function from matplotlib.pyplot library plots bar graph for given values of x and y.

A : plot()

B : draw()

C : bar()

D : linedraw()

Q.no 35. -----is not one of the key data science skill.

A : Statistics

B : Machine Learning

C : Data Visualization

D : software tester

Q.no 36. In matplotlib ----- function groups smaller axes that can exist togather within a single figure.

A : subplot()

B : divide_figure()

C : add_fig()

D : group_fig()

Q.no 37. ----- function is used to save an array as in image file.

A : matplotlib.pyplot.image()

B : matplotlib.pyplot.imread()

C : matplotlib.pyplot.imwrite()

D : matplotlib.pyplot.imsave()

Q.no 38. Entropy is a measure of the randomness in the information being processed.

A : Entropy

B : Support

C : Confidence

D : lift

Q.no 39. ----- function used to add two numpy arrays elementwise.

A : numpy.add(x1,x2)

B : numpy.mod(x1,x2)

C : numpy.true_divide(x1,x2)

D : numpy.addition(x1,x2)

Q.no 40. In this type of clustering each data type either belongs to a cluster completely or not.

A : Hard clustering

B : Soft Clustering

C : Medium clustering

D : Simple clustering

Q.no 41. The statement subplot(4,3,5) will divide figure into ----- and specify plotting sholud be done on plot number-----

A : 4 x 3, 5

B : 3x 4, 5

C : 3 x 5, 4

D : 5x 3, 4

Q.no 42. Select 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 : Analysed data is original source of data.

Q.no 43. Which function from numpy used to return the truncated value of the input elementwise?

A : round()

B : trunc()

C : del()

D : remove_decimal()

Q.no 44. Which function returns an ndarray object that contains the numbers that are evenly spaced on a log scale.

A : numpy.logspace()

B : numpy.log()

C : numpy.fill()

D : numpy.random()

Q.no 45. Which of the following statement will create an axes at the top right corner of the current figure

A : subplot(2,3,3)

B : subplot(2,3,2)

C : subplot(2,3,4)

D : subplot(2,3,5)

Q.no 46. ----- function is used to display an image through an external viewer in scipy.

A : display()

B : imread()

C : imshow()

D : show()

Q.no 47. To save a figure into a file we can use ----- method in the figure class of matplotlib.pyplot.

A : save()

B : save_fig()

C : Figure()

D : save_image()

Q.no 48. The ----- argument of merge function while merging two dataframes specifies which keys are to be included in the resulting dataframe.

A : right

B : on

C : sort

D : how

Q.no 49. ----- function performs the custom operations for the entire dataframe.

A : function()

B : surutine()

C : rutine()

D : pipe()

Q.no 50. ----- is basically extracting particular set of elements from an array.

A : Slicing

B : indexing

C : sorting

D : broadcasting

Q.no 51. To reach to the final point and to make prediction , decision trees must be traversed from -----

A : Top - to - bottom

B : Bottom- to - Top

C : Left- to Right

D : Right - to - Left

Q.no 52. Which of the following function is not used to iterate over the rows of the DataFrame.

A : iteritems()

B : iterrows()

C : itertuples()

D : iterpanel()

Q.no 53. Which of the following machine learning algorithm is used for market basket analysis means to analyze the association of purchased items in a single basket or single purchase.

A : Decision tree

B : Association Rule Mining

C : Clustering

D : Support vector machine

Q.no 54. Which of the following function is used to split a figure into nrow*ncols sub-axes.

A : plot()

B : draw()

C : bar()

D : subplot()

Q.no 55. In matplotlib ----- is container class for figure instance.

A : Axes

B : Canvas

C : Figure

D : FigureCanvas

Q.no 56. Which of the following algorithm is used in Economics, Finance, Biology etc, to model relationships between parameters of interests.

A : Regression

B : Decision Trees

C : Clustering

D : Naïve bays

Q.no 57. In regression the dependent variable is also called as -----

A : Regression

B : Continuous

C : Regressand

D : Independent

Q.no 58. ----- analysis is a set of statistical processes for estimating the relationships among dependent and independent variables.

A : Regression

B : Decision tree

C : KNN

D : None of These

Q.no 59. ----- algorithm models a series of logical If-Then- Else decision statements, there is no underlying assumption of a linear or non-linear relationship between the input variables and response variables.

A : Regression

B : Decision Trees

C : Clustering

D : Naïve bays

Q.no 60. In unsupervised learning, scikit learn uses ----- method to infer properties of the data.

A : extract()

B : transform()

C : infer()

D : classify()

Answer for Question No 1. is a

Answer for Question No 2. is c

Answer for Question No 3. is a

Answer for Question No 4. is c

Answer for Question No 5. is c

Answer for Question No 6. is a

Answer for Question No 7. is b

Answer for Question No 8. is b

Answer for Question No 9. is a

Answer for Question No 10. is a

Answer for Question No 11. is b

Answer for Question No 12. is b

Answer for Question No 13. is b

Answer for Question No 14. is a

Answer for Question No 15. is a

Answer for Question No 16. is d

Answer for Question No 17. is d

Answer for Question No 18. is b

Answer for Question No 19. is a

Answer for Question No 20. is a

Answer for Question No 21. is c

Answer for Question No 22. is c

Answer for Question No 23. is b

Answer for Question No 24. is a

Answer for Question No 25. is c

Answer for Question No 26. is c

Answer for Question No 27. is a

Answer for Question No 28. is a

Answer for Question No 29. is b

Answer for Question No 30. is d

Answer for Question No 31. is c

Answer for Question No 32. is a

Answer for Question No 33. is b

Answer for Question No 34. is c

Answer for Question No 35. is d

Answer for Question No 36. is a

Answer for Question No 37. is d

Answer for Question No 38. is a

Answer for Question No 39. is a

Answer for Question No 40. is a

Answer for Question No 41. is a

Answer for Question No 42. is a

Answer for Question No 43. is b

Answer for Question No 44. is a

Answer for Question No 45. is a

Answer for Question No 46. is c

Answer for Question No 47. is b

Answer for Question No 48. is d

Answer for Question No 49. is d

Answer for Question No 50. is a

Answer for Question No 51. is a

Answer for Question No 52. is d

Answer for Question No 53. is b

Answer for Question No 54. is d

Answer for Question No 55. is d

Answer for Question No 56. is a

Answer for Question No 57. is c

Answer for Question No 58. is a

Answer for Question No 59. is b

Answer for Question No 60. is b

Total number of questions : 60

13329_DATA ANALYTICS

Time : 1hr

Max Marks : 50

N.B

- 1) All questions are Multiple Choice Questions having single correct option.
 - 2) Attempt any 50 questions out of 60.
 - 3) Use of calculator is allowed.
 - 4) Each question carries 1 Mark.
 - 5) Specially abled students are allowed 20 minutes extra for examination.
 - 6) Do not use pencils to darken answer.
 - 7) Use only black/blue ball point pen to darken the appropriate circle.
 - 8) No change will be allowed once the answer is marked on OMR Sheet.
 - 9) Rough work shall not be done on OMR sheet or on question paper.
 - 10) Darken ONLY ONE CIRCLE for each answer.
-

Q.no 1. Naïve Bayes is a classification technique based on -----

A : Bayes Theorem

B : Pythagorous Theorum

C : Least square method

D : mean square method

Q.no 2. ----- function is used to plot a histogram using matplotlib library

A : hist()

B : bar()

C : pie()

D : scatter()

Q.no 3. ----- rule mining is a technique to identify underlying relations between different items.

A : Classification

B : Regression

C : Clustering

D : Association

Q.no 4. Probability always lies between ---- and ----

A : 0 and 1

B : -1 and +1

C : -1 and 0

D : 0 and infinite

Q.no 5. To import data from excel file into a dataframe ----- function is provided by pandas package.

A : read_csv()

B : read_file()

C : read()

D : read_excel()

Q.no 6. In numpy array , array indices always starts from -----

A : 1

B : -1

C : 0

D : 2

Q.no 7. Email data is an example of -----

A : Structured data

B : Un-Structured data

C : Semi-Structured data

D : Scattered

Q.no 8. ----- function used to get positive square root of an numpy array elementwise.

A : `numpy.sqrt(x1)`

B : `numpy.mod(x1)`

C : `numpy.square(x1)`

D : `numpy.find(x1,2)`

Q.no 9. In ----- learning the training is controlled by an external supervisor or teacher.

A : Un- Supervised

B : Supervised

C : semi-supervised

D : group

Q.no 10. For multidimensional visualization ----- are used.

A : pie charts

B : Bar charts

C : Andrews curves

D : Scatter plots

Q.no 11. The ----- algorithm is based on the fact that the algorithm uses prior knowledge to find frequent item set.

A : Clustering

B : Regression

C : Naïve Bayes

D : Apriori

Q.no 12. To import data from csv file into a dataframe ----- function is provided by pandas package.

A : `read_csv()`

B : read_file()

C : csv_read()

D : Frrom_csv()

Q.no 13. The ----- function creates a 2-D array with all values 1.

A : numpy.Ones()

B : numpy.zeros()

C : numpy.eye()

D : numpy.empty()

Q.no 14. K- nearest neighbors algorithm is based on ----- learning

A : Un- Supervised

B : Supervised

C : Association

D : correlation

Q.no 15. In support vector machines if input features are 2 then the decision boundries or hyperplane is -----.

A : 2-D plane

B : 3-D plane

C : Line

D : point

Q.no 16. ----- submodule of scipy is dedicated to image processing.

A : ndarray

B : spatial

C : ndimage

D : special

Q.no 17. ----- uses a tree structure to specify sequences of decisions and consequences.

A : Regression

B : Decision trees

C : KNN

D : SVM

Q.no 18. Numpy support this function to find trigonometric sine elementwise .

A : numpy.sin()

B : numpy.cosine()

C : numpy.tangent()

D : numpy.rad2sin(x1)

Q.no 19. The procedure to organize items of a given collection into groups based on some similar features called as -----

A : Regression

B : Clustering

C : Ddecion Trees

D : Association

Q.no 20. matplotlib.pyplot.imread() function is used to -----

A : save image

B : read image

C : copy image

D : show image

Q.no 21. ----- models search the data space for areas of varied density of data points in the data space.

A : Connectivity models

B : Centroid models

C : Distribution models

D : Density models

Q.no 22. Pandas provide ----- method in order to get purly integer based indexing.

A : iloc()

B : loc()

C : ix()

D : xloc()

Q.no 23. To rotate an image ----- function is used from scipy library.

A : rotation()

B : scipy.move()

C : scipy.ndimage.rotate()

D : scipy.flip()

Q.no 24. ----- is unsupervised machine learning technique.

A : KNN

B : Support Vector Machines

C : Decision trees

D : Cluster analysis

Q.no 25. -----is not one of the key data science skill.

A : Statistics

B : Machine Learning

C : Data Visualization

D : software tester

Q.no 26. The number of iterations in apriori -----

A : increases with the size of the data

B : decreases with the increase in size of the data

C : increases with the size of the maximum frequent set

D : decreases with increase in size of the maximum frequent set

Q.no 27. ----- regression finds a relationship between one or more features (independent variables) and a continuous variables (dependent variable).

A : Non-linear

B : Linear

C : Both of these

D : None of These

Q.no 28. -----is a flow-chart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and leaf nodes represent classes or class distributions.

A : Decision tree

B : Association Rule Mining

C : Clustering

D : Support vector machines

Q.no 29. Which of the following is not used for 2-D Visualisation?

A : pie charts

B : Bar charts

C : Andrews curves

D : Scatter plots

Q.no 30. Support(B) =

A : (Transacions containing (B)) / (Total Transactions)

B : (Transacions containing (B)) / 100

C : (Total Transactions) / (Transacions containing (B))

D : 100/ (Transacions containing (B))

Q.no 31. In decision trees leaf node denotes a -----

A : class distribution

B : test on an attribute

C : outcome of the test

D : class labels

Q.no 32. Which of the following is used as attribute selection measure in decision tree algorithms?

A : Information Gain

B : Posterior probability

C : Prior probability

D : Support

Q.no 33. A ----- is a supervised machine learning algorithm which relies on the assumption of feature independent to classify input data.

A : Clustering

B : Regression

C : Naïve Bayes

D : Apriori

Q.no 34. ----- function used to get arrays elementwise remainder of division

A : numpy.divide(x1,x2)

B : numpy.mod(x1,x2)

C : numpy.true_divide(x1,x2)

D : numpy.reminder(x1,x2)

Q.no 35. In this type of algorithms inputs are provided but not the desired output.

A : Cluster analysis

B : Support Vector Machines

C : Decision trees

D : Naïve bays

Q.no 36. ----- is an indication of how often the rule has been found to be true in association rule mining.

A : Confidence

B : Support

C : Lift

D : None of These

Q.no 37. ----- function from matplotlib.pyplot library plots bar graph for given values of x and y.

A : plot()

B : draw()

C : bar()

D : linedraw()

Q.no 38. To set x Axis lable of a figure----- function is used

A : set_title()

B : set_lable()

C : set_xlabel()

D : get_xlabel()

Q.no 39. What is the use of following function? Plt.xlabel("Total Marks")

A : Gives label to X-Axis

B : Gives label to Y-Axis

C : Gives title to figure

D : Add text to figure

Q.no 40. In SciPy ----- submodule is dedicated to image processing.

A : ndimage

B : ndarray

C : signal

D : io

Q.no 41. Apriori algorithm uses breadth first search and -----structure to count candidate item sets efficiently.

A : Decision tree

B : Hash tree

C : Red-Black Tree

D : AVL Tree

Q.no 42. Which of the following task is not performed by Data Scientist.

A : Define the question

B : Create reproducible code

C : Challenge results

D : Staff Recruitement

Q.no 43. To reach to the final point and to make prediction , decision trees must be traversed from -----

A : Top - to - bottom

B : Bottom- to - Top

C : Left- to Right

D : Right - to - Left

Q.no 44. Which of the following statement will create an axes at the top right corner of the current figure

A : subplot(2,3,3)

B : subplot(2,3,2)

C : subplot(2,3,4)

D : subplot(2,3,5)

Q.no 45. In regression the independent variable is also called as -----

A : Regressor

B : Continuous

C : Regressand

D : Estimated

Q.no 46. In unsupervised learning, scikit learn uses ----- method to infer properties of the data.

A : extract()

B : transform()

C : infer()

D : classify()

Q.no 47. Select 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 : Analysed data is original source of data.

Q.no 48. When there is no impact on one variable when increase or decrease on other variable then it is -----

A : Perfect correlation

B : No Correlation

C : Positive Correlation

D : Negative Correlation

Q.no 49. For testing accuracy of a machine learning algorithm whole data set should be devided into trainin and testing datasets. Which of the following is good proportion for train-test splitting?

A : Train- 70%, Test - 30%

B : Train- 50%, Test - 50%

C : Train- 30%, Test - 70%

D : Train- 100%, Test - 00%

Q.no 50. ----- analysis is a set of statistical processes for estimating the relationships among dependent and independent variables.

A : Regression

B : Decision tree

C : KNN

D : None of These

Q.no 51. Plot_number parameter from subplot() function can range from 1 to -----

A : nrows*ncols

B : max

C : nrows

D : ncols

Q.no 52. ----- algorithm models a series of logical If-Then- Else decision statements, there is no underlying assumption of a linear or non-linear relationship between the input variables and response variables.

A : Regression

B : Decision Trees

C : Clustering

D : Naïve bays

Q.no 53. ----- function from scipy is used to calculate the distance between all pairs of points in a given set.

A : scipy.spatial.distance()

B : scipy.spatial.distance.measure()

C : scipy.spatial.distance.cdist()

D : distance(x1,y1)

Q.no 54. In this type of clustering instead of putting each data point into a separate cluster a probability or likelihood of that data point to be in those clusters is assigned.

A : Hard clustering

B : Soft Clustering

C : Medium clustering

D : Simple clustring

Q.no 55. In regression the dependent variable is also called as -----

A : Regression

B : Continuous

C : Regressand

D : Independent

Q.no 56. The ----- argument of merge function while merging two dataframes specifies which keys are to be included in the resulting dataframe.

A : right

B : on

C : sort

D : how

Q.no 57. While plotting using matplotlib.pyplot A function call similar to subplot(2,3,4) is

A : subplot(234)

B : subplot(243)

C : subplot(324)

D : subplot(4)

Q.no 58. Catalog design is complex process where the selection of items in a business's catalog are often designed to complement each other so that buying one item will lead to buying of another. So these items are often complements or very related. Which algorithm

A : Decision tree

B : Association Rule Mining

C : Clustering

D : Support vector machine

Q.no 59. Which of the following function is used to split a figure into nrows*ncols sub-axes.

A : plot()

B : draw()

C : bar()

D : subplot()

Q.no 60. To save a figure into a file we can use ----- method in the figure class of matplotlib.pyplot.

A : save()

B : save_fig()

C : Figure()

D : save_image()

Answer for Question No 1. is a

Answer for Question No 2. is a

Answer for Question No 3. is d

Answer for Question No 4. is a

Answer for Question No 5. is d

Answer for Question No 6. is c

Answer for Question No 7. is b

Answer for Question No 8. is a

Answer for Question No 9. is b

Answer for Question No 10. is c

Answer for Question No 11. is d

Answer for Question No 12. is a

Answer for Question No 13. is a

Answer for Question No 14. is b

Answer for Question No 15. is c

Answer for Question No 16. is c

Answer for Question No 17. is b

Answer for Question No 18. is a

Answer for Question No 19. is b

Answer for Question No 20. is b

Answer for Question No 21. is d

Answer for Question No 22. is a

Answer for Question No 23. is c

Answer for Question No 24. is d

Answer for Question No 25. is d

Answer for Question No 26. is c

Answer for Question No 27. is b

Answer for Question No 28. is a

Answer for Question No 29. is c

Answer for Question No 30. is a

Answer for Question No 31. is c

Answer for Question No 32. is a

Answer for Question No 33. is c

Answer for Question No 34. is b

Answer for Question No 35. is a

Answer for Question No 36. is a

Answer for Question No 37. is c

Answer for Question No 38. is c

Answer for Question No 39. is a

Answer for Question No 40. is a

Answer for Question No 41. is b

Answer for Question No 42. is d

Answer for Question No 43. is a

Answer for Question No 44. is a

Answer for Question No 45. is a

Answer for Question No 46. is b

Answer for Question No 47. is a

Answer for Question No 48. is b

Answer for Question No 49. is a

Answer for Question No 50. is a

Answer for Question No 51. is a

Answer for Question No 52. is b

Answer for Question No 53. is c

Answer for Question No 54. is b

Answer for Question No 55. is c

Answer for Question No 56. is d

Answer for Question No 57. is a

Answer for Question No 58. is b

Answer for Question No 59. is d

Answer for Question No 60. is b

Seat No -

Total number of questions : 60

13329_DATA ANALYTICS

Time : 1hr

Max Marks : 50

N.B

- 1) All questions are Multiple Choice Questions having single correct option.
 - 2) Attempt any 50 questions out of 60.
 - 3) Use of calculator is allowed.
 - 4) Each question carries 1 Mark.
 - 5) Specially abled students are allowed 20 minutes extra for examination.
 - 6) Do not use pencils to darken answer.
 - 7) Use only black/blue ball point pen to darken the appropriate circle.
 - 8) No change will be allowed once the answer is marked on OMR Sheet.
 - 9) Rough work shall not be done on OMR sheet or on question paper.
 - 10) Darken ONLY ONE CIRCLE for each answer.
-

Q.no 1. Correlation coefficient values lies between---- and ---

A : -1 and +1

B : -1 and 0

C : 0 and 1

D : 0 and infinite

Q.no 2. -----type of analytics describes what happened in past

A : Descriptive

B : Prescriptive

C : Predictive

D : Probability

Q.no 3. In statistics, a population consists of -----

A : All People living in a country.

B : All People living in the city.

C : All subjects or objects whose characteristics are being studied.

D : Part of whole dataset

Q.no 4. SQL record is an example of -----

A : Structured data

B : Un-Structured data

C : Semi-Structured data

D : Scattered

Q.no 5. -----function reads an image from a file as an array.

A : imsave()

B : imread()

C : read()

D : None of these

Q.no 6. Find odd one out from the following :

A : KNN

B : Naïve Bayes

C : Decision Trees

D : Cluster analysis

Q.no 7. The ----- algorithm is based on the fact that the algorithm uses prior knowledge to find frequent item set.

A : Clustering

B : Regression

C : Naïve Bayes

D : Apriori

Q.no 8. Pin code of a city is an example of -----

A : Structured data

B : Un-Structured data

C : Semi-Structured data

D : Scattered

Q.no 9. matplotlib.pyplot.imread() function is used to -----

A : save image

B : read image

C : copy image

D : show image

Q.no 10. Choose correct option for machine generated unstructured data.

A : Website data

B : YouTube data

C : Text File data

D : Sensor data

Q.no 11. Which function is used to give title for the axes.

A : plt.title()

B : plt.xlabel()

C : plt.ylabel()

D : plt.xscale()

Q.no 12. Which of the following is measure used in decision trees while selecting splitting criteria that partitions data into the best possible manner.

A : Information Gain

B : Probability

C : Regression

D : Association

Q.no 13. ----- means part of population chosen for participation in the study

A : Population

B : Sample

C : Association

D : Correlation

Q.no 14. ----- is an example of human generated unstructured data.

A : YouTube data

B : Satellite data

C : Sensor data

D : Seismic imagery data

Q.no 15. ----- function is used to save image into an ndarray.

A : imsave()

B : imread()

C : save()

D : isave()

Q.no 16. ----- chart is a circular plot divides into slices to show numerical proportion.

A : Bar

B : Line

C : Scatter

D : Pie

Q.no 17. ----- answers the question "What will happen in future?"

A : Descriptive

B : Prescriptive

C : Predictive

D : Probability

Q.no 18. ----- method is dataframe reads first n rows from dataframe

A : head(n)

B : tail(n)

C : first(n)

D : start(n)

Q.no 19. ----- refers to the graphical representation of information and data.

A : Data Visualization

B : Data mining

C : Data warehousing

D : Data Structures

Q.no 20. ----- is a general purpose array-processing package provides a high performance multi-dimentional array object and tools for working with these arrays.

A : NumPy

B : SciPy

C : sklearn

D : None of these

Q.no 21. ----- is uses a tree structure to specify sequence of decisions and consequences.

A : KNN

B : NAïve Bayes

C : Regression

D : Decision Tree

Q.no 22. Which statement will create 5 x 5 array filled with all values 1

A : x=numpy.ones((5,5))

B : x=numpy.ones(5)

C : x=numpy.zeros((5,5))

D : x=numpy.eye((5,5))

Q.no 23. In matplotlib library ----- module supports basic image loading, rescaling and display operations.

A : picture

B : image

C : pyplot

D : sympy

Q.no 24. ----- function used to get arrays elementwise remainder of division

A : numpy.divide(x1,x2)

B : numpy.mod(x1,x2)

C : numpy.true_divide(x1,x2)

D : numpy.reminder(x1,x2)

Q.no 25. In ----- the x-axes are grouped into bins and each bin will be treated as a category.

A : Bar

B : Line

C : Scatter

D : Histogram

Q.no 26. ----- is most important language for Data Science.

A : Java

B : Ruby

C : R

D : None of these

Q.no 27. The ---- algorithm is the simplest machine learning algorithm, which building the model consists only of storing the training dataset. To make a prediction for a new data point, the algorithm finds the closest data points in the training dataset i.e its

A : Apriori

B : K-Nearest Neighbors

C : K-Means

D : Decision Trees

Q.no 28. From matplotlib----- module is used for plotting various plots.

A : Scilearn

B : Pyplot

C : Scilab

D : Matlab

Q.no 29. Among the following clustering algorithm types in which of the following type the notion of similarity is derived by the closeness of a data point to the centroid of the clusters.

A : Connectivity models

B : Centroid models

C : Distribution models

D : Density models

Q.no 30. ----- is a form of supervised learning algorithm which is used in mail service providers like Gmail, yahoo, etc. to classify a new mail as spam or not spam.

A : Classification

B : Regression

C : Clustering

D : Naïve bays

Q.no 31. The number of iterations in apriori -----

A : increases with the size of the data

B : decreases with the increase in size of the data

C : increases with the size of the maximum frequent set

D : decreases with increase in size of the maximum frequent set

Q.no 32. In this type of algorithms inputs are provided but not the desired output.

A : Cluster analysis

B : Support Vector Machines

C : Decision trees

D : Naïve bays

Q.no 33. The objective of ----- algorithm is to find a hyperplane in an N-dimensional space that distinctly classifies the data points.

A : KNN

B : Support Vector Machines

C : Regression

D : Decision Tree

Q.no 34. Which of the following is used as attribute selection measure in decision tree algorithms?

A : Information Gain

B : Posterior probability

C : Prior probability

D : Support

Q.no 35. ----- analysis finds the reasons behind success or failure in past

A : Descriptive

B : Prescriptive

C : Predictive

D : Probability

Q.no 36. A -----graph is a circular plot, divided into slices to show numerical proportions.

A : Bar

B : Scatter

C : pie

D : line

Q.no 37. Support(B) =

A : (Transacions containing (B)) / (Total Transactions)

B : (Transacions containing (B)) / 100

C : (Total Transactions) / (Transacions containing (B))

D : 100/ (Transacions containing (B))

Q.no 38. -----is not one of the key data science skill.

A : Statistics

B : Machine Learning

C : Data Visualization

D : software tester

Q.no 39. ----- is an indication of how frequently the itemset appears in the dataset in association rule mining.

A : Confidence

B : Support

C : Lift

D : None of These

Q.no 40. When data are collected in a statistical study for only a portion or subset of all elements of interest we are using

A : Sample

B : Parameter

C : Population

D : Probability

Q.no 41. In Data science project data acquisition step involves-----

A : Acquiring data from various sources.

B : Selecting dataset

C : Data preprocessing

D : Data modeling

Q.no 42. In unsupervised learning, scikit learn uses ----- method to infer properties of the data.

A : extract()

B : transform()

C : infer()

D : classify()

Q.no 43. The -- ---- is characterized by a bell shaped curve and area under curve represents probabilities

A : Normal Distribution

B : Binomial Distribution

C : Poission Distribution

D : Probability

Q.no 44. ----- algorithm models a series of logical If-Then- Else decision statements, there is no underlying assumption of a linear or non-linear relationship between the input variables and response variables.

A : Regression

B : Decision Trees

C : Clustering

D : Naïve bays

Q.no 45. Which function returns an ndarray object that contains the numbers that are evenly spaced on a log scale.

A : numpy.logspace()

B : numpy.log()

C : numpy.fill()

D : numpy.random()

Q.no 46. To reach to the final point and to make prediction , decision trees must be traversed from -----

A : Top - to - bottom

B : Bottom- to - Top

C : Left- to Right

D : Right - to - Left

Q.no 47. ----- is an unsupervised algorithm used for frequent itemset mining.

A : Apriori

B : Support Vector Machines

C : Decision trees

D : Cluster analysis

Q.no 48. Which function from numpy used to return the truncated value of the input elementwise?

A : round()

B : trunc()

C : del()

D : remove_decimal()

Q.no 49. The strength (degree) of the correlation between a set of independent variables X and a dependent variable Y is measured by-----

A : Coefficient of Correlation

B : Coefficient of Determination

C : Standard error of estimate

D : Probability

Q.no 50. Which of the following function is not used to iterate over the rows of the DataFrame.

A : iteritems()

B : iterrows()

C : itertuples()

D : iterpanel()

Q.no 51. Which of the following statement will create an axes at the top right corner of the current figure

A : subplot(2,3,3)

B : subplot(2,3,2)

C : subplot(2,3,4)

D : subplot(2,3,5)

Q.no 52. It is a measure of disorder or purity or unpredictability or uncertainty.

A : Entropy

B : Support

C : Confidence

D : lift

Q.no 53. ----- function performs the custom operations for the entire dataframe.

A : function()

B : surutine()

C : rutine()

D : pipe()

Q.no 54. The ----- argument of merge function while merging two dataframes specifies which keys are to be included in the resulting dataframe.

A : right

B : on

C : sort

D : how

Q.no 55. Which of the following machine learning algorithm is used for maret basket analysis means to analyze the association of purchased items in a single basket or single purchase.

A : Decision tree

B : Association Rule Mining

C : Clustering

D : Support vector machine

Q.no 56. ----- analysis is a set of statistical processes for estimating the relationships among dependent and independent variables.

A : Regression

B : Decision tree

C : KNN

D : None of These

Q.no 57. To save a figure into a file we can use ----- method in the figure class of matplotlib.pyplot.

A : save()

B : save_fig()

C : Figure()

D : save_image()

Q.no 58. Which of the following algorithm is used in Economics, Finance, Biology etc, to model relationships between parameters of interests.

A : Regression

B : Decision Trees

C : Clustering

D : Naïve bays

Q.no 59. While plotting using matplotlib.pyplot A function call similar to subplot(2,3,4) is

A : subplot(234)

B : subplot(243)

C : subplot(324)

D : subplot(4)

Q.no 60. Apriori algorithm uses breadth first search and -----structure to count candidate item sets efficiently.

A : Decision tree

B : Hash tree

C : Red-Black Tree

D : AVL Tree

Answer for Question No 1. is a

Answer for Question No 2. is a

Answer for Question No 3. is c

Answer for Question No 4. is a

Answer for Question No 5. is b

Answer for Question No 6. is d

Answer for Question No 7. is d

Answer for Question No 8. is a

Answer for Question No 9. is b

Answer for Question No 10. is d

Answer for Question No 11. is a

Answer for Question No 12. is a

Answer for Question No 13. is b

Answer for Question No 14. is a

Answer for Question No 15. is a

Answer for Question No 16. is d

Answer for Question No 17. is c

Answer for Question No 18. is a

Answer for Question No 19. is a

Answer for Question No 20. is a

Answer for Question No 21. is d

Answer for Question No 22. is a

Answer for Question No 23. is b

Answer for Question No 24. is b

Answer for Question No 25. is d

Answer for Question No 26. is c

Answer for Question No 27. is b

Answer for Question No 28. is b

Answer for Question No 29. is b

Answer for Question No 30. is a

Answer for Question No 31. is c

Answer for Question No 32. is a

Answer for Question No 33. is b

Answer for Question No 34. is a

Answer for Question No 35. is a

Answer for Question No 36. is c

Answer for Question No 37. is a

Answer for Question No 38. is d

Answer for Question No 39. is b

Answer for Question No 40. is a

Answer for Question No 41. is a

Answer for Question No 42. is b

Answer for Question No 43. is a

Answer for Question No 44. is b

Answer for Question No 45. is a

Answer for Question No 46. is a

Answer for Question No 47. is a

Answer for Question No 48. is b

Answer for Question No 49. is a

Answer for Question No 50. is d

Answer for Question No 51. is a

Answer for Question No 52. is a

Answer for Question No 53. is d

Answer for Question No 54. is d

Answer for Question No 55. is b

Answer for Question No 56. is a

Answer for Question No 57. is b

Answer for Question No 58. is a

Answer for Question No 59. is a

Answer for Question No 60. is b

Seat No -

Total number of questions : 60

13329_DATA ANALYTICS

Time : 1hr

Max Marks : 50

N.B

- 1) All questions are Multiple Choice Questions having single correct option.
- 2) Attempt any 50 questions out of 60.
- 3) Use of calculator is allowed.
- 4) Each question carries 1 Mark.
- 5) Specially abled students are allowed 20 minutes extra for examination.
- 6) Do not use pencils to darken answer.
- 7) Use only black/blue ball point pen to darken the appropriate circle.
- 8) No change will be allowed once the answer is marked on OMR Sheet.
- 9) Rough work shall not be done on OMR sheet or on question paper.
- 10) Darken ONLY ONE CIRCLE for each answer.

Q.no 1. ----- analysis estimates the relationship between single dependent variable and single independent variable

A : Simple Regression

B : Multiple regression

C : Correlation

D : Probability

Q.no 2. Find odd one out from the following :

A : KNN

B : NAïve Bayes

C : Decision Trees

D : Cluster analysis

Q.no 3. ----- chart is a circular plot divides into slices to show numerical proportion.

A : Bar

B : Line

C : Scatter

D : Pie

Q.no 4. ----- type of plots show all individual data points without connected with lines.

A : Bar

B : Line

C : Scatter

D : Histogram

Q.no 5. Which of the following is NOT supervised learning?

A : PCA

B : Decision Tree

C : Linear Regression

D : Naive Bayesian

Q.no 6. Probability always lies between ---- and ----

A : 0 and 1

B : -1 and +1

C : -1 and 0

D : 0 and infinite

Q.no 7. In numpy array , array indices always starts from -----

A : 1

B : -1

C : 0

D : 2

Q.no 8. To import data from excel file into a dataframe ----- function is provided by pandas package.

A : read_csv()

B : read_file()

C : read()

D : read_excel()

Q.no 9. ----- plot displays information as series of data points connected by straight lines.

A : Bar

B : Line

C : Scatter

D : Histogram

Q.no 10. Which of the following is not a raster image file format?

A : PNG

B : JPG

C : BMP

D : PDF

Q.no 11. Naïve Bayes is a classification technique based on -----

A : Bayes Theorem

B : Pythagorous Theorum

C : Least square method

D : mean square method

Q.no 12. ---- is an technique to learn from examples and experience, without being explicitly programmed.

A : Machine Learning

B : Software Testing

C : Computer Science

D : Data mining

Q.no 13. ----- library is built on the top of Numpy, SciPy and Matplotlib

A : Sympy

B : Scikit

C : Pandas

D : Numpy

Q.no 14. ----- function is used to save image into an ndarray.

A : imsave()

B : imread()

C : save()

D : isave()

Q.no 15. For multidimensional visualization ----- are used.

A : pie charts

B : Bar charts

C : Andrews curves

D : Scatter plots

Q.no 16. ----- library from python provides efficient versions of a large number of machine learning algorithms.

A : Pandas

B : Numpy

C : Scikit-Learn

D : image

Q.no 17. In statistics, a population consists of -----

A : All People living in a country.

B : All People living in the city.

C : All subjects or objects whose characteristics are being studied.

D : Part of whole dataset

Q.no 18. Which library from python is used for implementing machine learning algorithms?

A : Scikit-Learn

B : Pandas

C : Matplotlib

D : Numpy

Q.no 19. SQL record is an example of -----

A : Structured data

B : Un-Structured data

C : Semi-Structured data

D : Scattered

Q.no 20. ----- is about developing code to enable the machine to learn to perform tasks and its basic principle is the automatic modeling of underlying that have generated the collected data.

A : Data Science

B : Data Analytics

C : Data Warehousing

D : Data mining

Q.no 21. ----- is the measure of the likeihood that an event will occure in a random experiment

A : Probability

B : Correlation

C : Regression

D : Sample

Q.no 22. Entropy is a measure of the randomness in the information being processed.

A : Entropy

B : Support

C : Confidence

D : lift

Q.no 23. In head() / tail() functions of dataframe the default number of elements to display is -----

A : 3

B : 5

C : 1

D : 10

Q.no 24. In SciPy ----- submodule is dedicated to image processing.

A : ndimage

B : ndarray

C : signal

D : io

Q.no 25. ----- module from sklearn gathers popular unsupervised clustering algorithms.

A : sklearn.covariance

B : sklearn.base

C : sklearn.neighbors

D : sklearn.cluster

Q.no 26. ----- function used to get arrays elementwise remainder of division

A : numpy.divide(x1,x2)

B : numpy.mod(x1,x2)

C : numpy.true_divide(x1,x2)

D : numpy.reminder(x1,x2)

Q.no 27. Which of the following plots is not used for multidimensional visualization?

A : Andrrews Curves

B : Prallel Chart

C : Deviation Chart

D : Bar

Q.no 28. ----- searches for the linear optimal separating hyperplane for separation of the data using essential training tuples called support vectors

A : Decision tree

B : Association Rule Mining

C : Clustering

D : Support vector machines

Q.no 29. From matplotlib----- module is used for plotting various plots.

A : Scilearn

B : Pyplot

C : Scilab

D : Matlab

Q.no 30. In ----- the x-axes are grouped into bins and each bin will be treated as a category.

A : Bar

B : Line

C : Scatter

D : Histogram

Q.no 31. If X and Y are both independent of each other, then correlation coefficient is -----

A : 1

B : -1

C : 0

D : 2

Q.no 32. ----- is an indication of how often the rule has been found to be true in association rule mining.

A : Confidence

B : Support

C : Lift

D : None of These

Q.no 33. Among the following clustering algorithm types in which of the following type the notion of similarity is derived by the closeness of a data point to the centroid of the clusters.

A : Connectivity models

B : Centroid models

C : Distribution models

D : Density models

Q.no 34. The last element of ndarray is indexed by -----

A : 0

B : -1

C : 1

D : -2

Q.no 35. ----- changes the the arrangement of items form array so that shape of array changes while maintaining the same number of dimensions.

A : numpy. Reshape()

B : numpy. Empty()

C : numpy. Flatten()

D : numpy.ravel()

Q.no 36. Identify the machine generated unstructured data.

A : Website data

B : YouTube data

C : Text File data

D : Satellite imagery data

Q.no 37. ----- is unsupervised machine learning technique.

A : KNN

B : Support Vector Machines

C : Decision trees

D : Cluster analysis

Q.no 38. Support(B) =

A : (Transacions containing (B)) / (Total Transactions)

B : (Transacions containing (B)) / 100

C : (Total Transactions) / (Transacions containing (B))

D : 100/ (Transacions containing (B))

Q.no 39. ----- is an example of semi structured data

A : XML data

B : YouTube data

C : Text File data

D : Satellite imagery data

Q.no 40. In decision trees leaf node denotes a -----

A : class distribution

B : test on an attribute

C : outcome of the test

D : class labels

Q.no 41. Which of the following algorithm is used in Economics, Finance, Biology etc, to model relationships between parameters of interests.

A : Regression

B : Decision Trees

C : Clustering

D : Naïve bays

Q.no 42. In regression the dependent variable is also called as -----

A : Regression

B : Continuous

C : Regressand

D : Independent

Q.no 43. In regression the independent variable is also called as -----

A : Regressor

B : Continuous

C : Regressand

D : Estimated

Q.no 44. Which of the following function is not used to iterate over the rows of the DataFrame.

A : iteritems()

B : iterrows()

C : itertuples()

D : iterpanel()

Q.no 45. ----- analysis is a set of statistical processes for estimating the relationships among dependent and independent variables.

A : Regression

B : Decision tree

C : KNN

D : None of These

Q.no 46. In unsupervised learning, scikit learn uses ----- method to infer properties of the data.

A : extract()

B : transform()

C : infer()

D : classify()

Q.no 47. To reach to the final point and to make prediction , decision trees must be traversed from -----

A : Top - to - bottom

B : Bottom- to - Top

C : Left- to Right

D : Right - to - Left

Q.no 48. The -- ---- is characterized by a bell shaped curve and area under curve represents probabilities

A : Normal Distribution

B : Binomial Distribution

C : Poission Distribution

D : Probability

Q.no 49. Which of the following function is used to split a figure into nrow*ncols sub-axes.

A : plot()

B : draw()

C : bar()

D : subplot()

Q.no 50. In Data science project data acquisition step involves-----

A : Acquiring data from various sources.

B : Selecting dataset

C : Data preprocessing

D : Data modeling

Q.no 51. ----- function from scipy is used to calculate the distance between all pairs of points in a given set.

A : `scipy.spatial.distance()`

B : `scipy.spatial.distance.measure()`

C : `scipy.spatial.distance.cdist()`

D : `distance(x1,y1)`

Q.no 52. Which function returns an ndarray object that contains the numbers that are evenly spaced on a log scale.

A : `numpy.logspace()`

B : `numpy.log()`

C : `numpy.fill()`

D : `numpy.random()`

Q.no 53. In matplotlib ----- is container class for figure instance.

A : `Axes`

B : `Canvas`

C : `Figure`

D : `FigureCanvas`

Q.no 54. ----- machine learning algorithm used in cross marketing to work with other businesss that complement your own business but not to other competitors.

A : Decision tree

B : Association Rule Mining

C : Clustering

D : Support vector machine

Q.no 55. Select 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 : Analysed data is original source of data.

Q.no 56. It is a measure of disorder or purity or unpredictability or uncertainty.

A : Entropy

B : Support

C : Confidence

D : lift

Q.no 57. To determine basic salary of a employee when his qualification is given is a ----- problem

A : Correlation

B : Regression

C : Association

D : Qualitative

Q.no 58. The statement subplot(4,3,5) will divide figure into ----- and specify plotting sholud be done on plot number-----

A : 4 x 3, 5

B : 3x 4, 5

C : 3 x 5, 4

D : 5x 3, 4

Q.no 59. ----- algorithm models a series of logical If-Then- Else decision statements, there is no underlying assumption of a linear or non-linear relationship between the input variables and response variables.

A : Regression

B : Decision Trees

C : Clustering

D : Naïve bays

Q.no 60. ----- function is used to display an image through an external viewer in scipy.

A : display()

B : imread()

C : imshow()

D : show()

Answer for Question No 1. is a

Answer for Question No 2. is d

Answer for Question No 3. is d

Answer for Question No 4. is c

Answer for Question No 5. is a

Answer for Question No 6. is a

Answer for Question No 7. is c

Answer for Question No 8. is d

Answer for Question No 9. is b

Answer for Question No 10. is d

Answer for Question No 11. is a

Answer for Question No 12. is a

Answer for Question No 13. is b

Answer for Question No 14. is a

Answer for Question No 15. is c

Answer for Question No 16. is c

Answer for Question No 17. is c

Answer for Question No 18. is a

Answer for Question No 19. is a

Answer for Question No 20. is b

Answer for Question No 21. is a

Answer for Question No 22. is a

Answer for Question No 23. is b

Answer for Question No 24. is a

Answer for Question No 25. is d

Answer for Question No 26. is b

Answer for Question No 27. is d

Answer for Question No 28. is d

Answer for Question No 29. is b

Answer for Question No 30. is d

Answer for Question No 31. is b

Answer for Question No 32. is a

Answer for Question No 33. is b

Answer for Question No 34. is b

Answer for Question No 35. is a

Answer for Question No 36. is d

Answer for Question No 37. is d

Answer for Question No 38. is a

Answer for Question No 39. is a

Answer for Question No 40. is c

Answer for Question No 41. is a

Answer for Question No 42. is c

Answer for Question No 43. is a

Answer for Question No 44. is d

Answer for Question No 45. is a

Answer for Question No 46. is b

Answer for Question No 47. is a

Answer for Question No 48. is a

Answer for Question No 49. is d

Answer for Question No 50. is a

Answer for Question No 51. is c

Answer for Question No 52. is a

Answer for Question No 53. is d

Answer for Question No 54. is b

Answer for Question No 55. is a

Answer for Question No 56. is a

Answer for Question No 57. is b

Answer for Question No 58. is a

Answer for Question No 59. is b

Answer for Question No 60. is c

Data Analytics Mcq with Answers

1. Business intelligence (BI) is a broad category of application programs which includes _

1. Decision support
2. Data mining
3. OLAP
4. All of the mentioned

Show Answer

All of the mentioned

2. BI can catalyze a business's success in terms of _

1. Distinguish the products and services that drive revenues
2. Rank customers and locations based on profitability
3. Ranks customers and locations based on probability
4. All of the mentioned

Show Answer

All of the mentioned

3. Which of the following areas are affected by BI?

1. Revenue
2. CRM
3. Sales
4. All of the mentioned

Show Answer

CRM(Customer relationship management)

4. __ is a performance management tool that recapitulates an organization's performance from several standpoints on a single page

Data Analytics Mcq with Answers

1. Balanced Scorecard
2. Data Cube
3. Dashboard
4. All of the mentioned

Show Answer

Balanced Scorecard

5. __ is a system where operations like data extraction, transformation and loading operations are executed.

1. Data staging
2. Data integration
3. ETL
4. None of the mentioned

Show Answer

Data staging

6. _____ is a category of applications and technologies for presenting and analyzing corporate and external data.

1. Data warehouse
2. MIS
3. EIS
4. All of the mentioned

Show Answer

EIS(Enterprise Information System)

7. Which of the following is the process of basing an organization's actions and decisions on actual measured results of performance?

1. Institutional performance management

Data Analytics Mcq with Answers

2. Gap analysis
3. Slice and Dice
4. None of the mentioned

Show Answer

Institutional performance management

8. Which of the following does not form part of BI Stack in SQL Server?

1. SSRS
2. SSIS
3. SSAS
4. OBIEE

Show Answer

OBIEE

9. BI can catalyze a business's success in terms of _____

1. Distinguish the products and services that drive revenues
2. Rank customers and locations based on profitability
3. Ranks customers and locations based on probability
4. All of the mentioned

Show Answer

All of the mentioned

10. This is an approach to selling goods and services in which a prospect explicitly agrees in advance to receive marketing information

1. customer managed relationship
2. data mining
3. permission marketing
4. one-to-one marketing

Data Analytics Mcq with Answers

Show Answer

permission marketing

11. In an Internet context, this is the practice of tailoring Web pages to individual users' characteristics or preferences.

1. Web services
2. customer-facing
3. client/server
4. personalization

Show Answer

personalization

12. This is the processing of data about customers and their relationship with the enterprise in order to improve the enterprise's future sales and service and lower cost.

1. clickstream analysis
2. database marketing
3. customer relationship management
4. CRM analytics

Show Answer

CRM analytics

13. This is a broad category of applications and technologies for gathering, storing, analyzing, and providing access to data to help enterprise users make better business decisions.

1. best practice
2. data mart
3. business information warehouse
4. business intelligence

Data Analytics Mcq with Answers

Show Answer

business intelligence

14. This is a systematic approach to the gathering, consolidation, and processing of consumer data (both for customers and potential customers) that is maintained in a company's databases

1. database marketing
2. marketing encyclopedia
3. application integration
4. service oriented integration

Show Answer

database marketing

15. This is an arrangement in which a company outsources some or all of its customer relationship management functions to an application service provider (ASP).

1. spend management
2. supplier relationship management
3. hosted CRM
4. Customer Information Control System

Show Answer

hosted CRM

16. This is an XML-based metalanguage developed by the Business Process Management Initiative (BPMI) as a means of modeling business processes, much as XML is, itself, a metalanguage with the ability to model enterprise data.

1. BizTalk
2. BPML
3. e-biz

Data Analytics Mcq with Answers

4. ebXML b

Show Answer

BPML

17. This is a central point in an enterprise from which all customer contacts are managed.

1. contact center
2. help system
3. multichannel marketing
4. call center

Show Answer

contact center

18. This is the practice of dividing a customer base into groups of individuals that are similar in specific ways relevant to marketing, such as age, gender, interests, spending habits, and so on.

1. customer service chat
2. customer managed relationship
3. customer life cycle
4. customer segmentation

Show Answer

customer segmentation

19. In data mining, this is a technique used to predict future behavior and anticipate the consequences of change.

1. predictive technology
2. disaster recovery
3. phase change
4. predictive modeling

Data Analytics Mcq with Answers

Show Answer

predictive modeling

20. According to analysts, for what can traditional IT systems provide a foundation when they're integrated with big data technologies like Hadoop?"

1. Big data management and data mining
2. Data warehousing and business intelligence
3. Management of Hadoop clusters
4. Collecting and storing unstructured data

Show Answer

Big data management and data mining

data analytics mcq questions and answers

21. All of the following accurately describe Hadoop, EXCEPT:

1. Open source
2. Real-time
3. Java-based
4. Distributed computing approach

Show Answer

Real-time

22. ____has the world's largest Hadoop cluster

1. Apple
2. Datamatics
3. Facebook
4. None of the mentioned

Show Answer

Data Analytics Mcq with Answers

Facebook

23. What are the five V's of Big Data?

1. Volume
2. velocity
3. Variety
4. All of the above

Show Answer

All of the above

24. ____ hides the limitations of Java behind a powerful and concise Clojure API for Cascading.”

1. Scalding
2. Cascalog
3. Hcatalog
4. Hcalding

Show Answer

Cascalog

25. What are the main components of Big Data?

1. MapReduce
2. HDFS
3. YARN
4. All of these

Show Answer

All of these

26. What are the different features of Big Data Analytics?

1. Open-Source

Data Analytics Mcq with Answers

2. Scalability
3. Data Recovery
4. All the above

Show Answer

All the above

27. Define the Port Numbers for NameNode, Task Tracker and Job Tracker

1. NameNode
2. Task Tracker
3. Job Tracker
4. All of the above

Show Answer

All of the above

28. Facebook Tackles Big Data With ____ based on Hadoop

1. Project Prism
2. Prism
3. ProjectData
4. ProjectBid

Show Answer

Project Prism

29. What is a unit of data that flows through a Flume agent?

1. Record
2. Event
3. Row
4. Log

Data Analytics Mcq with Answers

Show Answer

Event

30. A feature F1 can take certain value: A, B, C, D, E, & F and represents grade of students from a college. Which of the following statement is true in the following case

1. Feature F1 is an example of nominal variable
2. Feature F1 is an example of ordinal variable
3. It doesn't belong to any of the above category
4. Both of these

Show Answer

Feature F1 is an example of ordinal variable

data analytics mcq with answers

31. Which of the following is an example of a deterministic algorithm?

1. PCA
2. K-Means
3. None of the above
4. all of the above

Show Answer

PCA

32. What is the entropy of the target variable?

1. $-(5/8 \log(5/8) + 3/8 \log(3/8))$
2. $5/8 \log(5/8) + 3/8 \log(3/8)$
3. $5/8 \log(5/8) + 3/8 \log(3/8)$
4. $5/8 \log(3/8) - 3/8 \log(5/8)$

Show Answer

Data Analytics Mcq with Answers

$-(5/8 \log(5/8) + 3/8 \log(3/8))$

33. Point out the correct statement.

1. OLAP is an umbrella term that refers to an assortment of software applications for analyzing an organization's raw data for intelligent decision making
2. Business intelligence equips enterprises to gain business advantage from data
3. BI makes an organization agile thereby giving it a lower edge in today's evolving market condition
4. None of the mentioned

Show Answer

Business intelligence equips enterprises to gain business advantage from data

34. BI can catalyze a business's success in terms of _____

1. Distinguish the products and services that drive revenues
2. Rank customers and locations based on profitability
3. Ranks customers and locations based on probability
4. All of the mentioned

Show Answer

All of the mentioned

data analytics multiple choice questions

35. Heuristic is

1. A set of databases from different vendors, possibly using different database paradigms
2. An approach to a problem that is not guaranteed to work but performs well in most cases
3. Information that is hidden in a database and that cannot be recovered by a simple SQL query.

Data Analytics Mcq with Answers

4. None of these

Show Answer

An approach to a problem that is not guaranteed to work but performs well in most cases

36. Heterogeneous databases referred to

1. A set of databases from different b vendors, possibly using different database paradigms
2. An approach to a problem that is not guaranteed to work but performs well in most cases.
3. Information that is hidden in a database and that cannot be recovered by a simple SQL query.
4. None of these

Show Answer

A set of databases from different b vendors, possibly using different database paradigms

1. Is it possible that Assignment of observations to clusters does not change between successive iterations in K-Means

1. Yes
2. No
3. Can't say
4. None of these

Show Answer

Yes

2. Which of the following can act as possible termination conditions in K-Means?

Data Analytics Mcq with Answers

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.
5. All of the above

Show Answer

All of the above

3. Which of the following clustering algorithms suffers from the problem of convergence at local optima?

1. K- Means clustering algorithm
2. Agglomerative clustering algorithm
3. Expectation-Maximization clustering algorithm
4. Diverse clustering algorithm
5. both a and c

Show Answer

both a and c

4. How can Clustering (Unsupervised Learning) be used to improve the accuracy of Linear Regression model (Supervised Learning):

1. Creating different models for different cluster groups.
2. Creating an input feature for cluster ids as an ordinal variable.
3. Creating an input feature for cluster centroids as a continuous variable.
4. Creating an input feature for cluster size as a continuous variable.
5. All of the above

Show Answer

Data Analytics Mcq with Answers

All of the above

5. What could be the possible reason(s) for producing two different dendrograms using agglomerative clustering algorithm for the same dataset? because

1. Proximity function used
2. of data points used
3. of variables used
4. All of the above

Show Answer

All of the above

6. 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
5. a, b and d

Show Answer

a, b and d

7. Which of the following is/are valid iterative strategy for treating missing values before clustering analysis?

1. Imputation with mean
2. Nearest Neighbor assignment
3. computation with Expectation
4. Maximization algorithm All of the above

Show Answer

Data Analytics Mcq with Answers

computation with Expectation

8. Feature scaling is an important step before applying K-Mean algorithm. What is reason behind this?

1. In distance calculation it will give the same weights for all features
2. You always get the same clusters. If you use or don't use feature scaling
3. In Manhattan distance it is an important step but in Euclidian it is not
4. None of these

Show Answer

In distance calculation it will give the same weights for all features

9. Which of the following method is used for finding optimal of cluster in K-Mean algorithm?

1. Elbow method
2. Manhattan method
3. Euclidian mehthod
4. All of the above

Show Answer

Elbow method

10. What is true about K-Mean Clustering?

1. K-means is extremely sensitive to cluster center initializations
2. Bad initialization can lead to Poor convergence speed
3. Bad initialization can lead to bad overall clustering
4. None of these

Show Answer

None of these

Data Analytics mcq sppu

Data Analytics Mcq with Answers

11. Which of the following can be applied to get good results for K-means algorithm corresponding to global minima?

1. Try to run algorithm for different centroid initialization
2. Adjust number of iterations
3. Find out the optimal number of clusters
4. All of the above

Show Answer

All of the above

12. If you are using Multinomial mixture models with the expectation-maximization algorithm for clustering a set of data points into two clusters, which of the assumptions are important:

1. All the data points follow two Gaussian distribution
2. All the data points follow n Gaussian distribution ($n > 2$)
3. All the data points follow two multinomial distribution
4. All the data points follow n multinomial distribution ($n > 2$)

Show Answer

All the data points follow two multinomial distribution

13. Which of the following is/are not true about Centroid based K-Means clustering algorithm and Distribution based expectation-maximization clustering algorithm:

1. Both starts with random initializations
2. Both are iterative algorithms
3. Both have strong assumptions that the data points must fulfill
4. Expectation maximization algorithm is a special case of K-Means

Show Answer

Expectation maximization algorithm is a special case of K-Means

Data Analytics Mcq with Answers

14. Which of the following is/are not true about DBSCAN clustering algorithm:

1. For data points to be in a cluster, they must be in a distance threshold to a core point
2. It has strong assumptions for the distribution of data points in dataspace
3. It has substantially high time complexity of order $O(n^3)$
4. It does not require prior knowledge of the no. of desired clusters
5. both b and c

Show Answer

both b and c

15. Which of the following are the high and low bounds for the existence of F-Score?

1. [0,1]
2. (0,1)
3. [-1,1]
4. None of the above

Show Answer

[0,1]

16. All of the following increase the width of a confidence interval except:

1. Increased confidence level
2. Increased variability
3. Increased sample size
4. Decreased sample size

Show Answer

Increased sample size

Data Analytics Mcq with Answers

17. The p-value in hypothesis testing represents which of the following:

Please select the best answer of those provided below

1. The probability of failing to reject the null hypothesis, given the observed results
2. The probability that the null hypothesis is true, given the observed results
3. The probability that the observed results are statistically significant, given that the null hypothesis is true
4. The probability of observing results as extreme or more extreme than currently observed, given that the null hypothesis is true

Show Answer

The probability of observing results as extreme or more extreme than currently observed, given that the null hypothesis is true

18. Assume that the difference between the observed, paired sample values is defined in the same manner and that the specified significance level is the same for both hypothesis tests. Using the same data, the statement that “a paired/dependent two sample t-test is equivalent to a one sample t-test on the paired differences, resulting in the same test statistic, same p-value, and same conclusion” is: Please select the best answer of those provided below.

1. Always True
2. Never True
3. Sometimes True
4. Not Enough Information

Show Answer

Always True

19. Green sea turtles have normally distributed weights, measured in kilograms, with a mean of 134.5 and a variance of 49.0. A particular green sea turtle’s weight has a z-score of -2.4. What is the weight of this green sea turtle? Round to the nearest whole number.

1. 17 kg

Data Analytics Mcq with Answers

- 2. 151 kg
- 3. 118 kg
- 4. 252 kg c

Show Answer

118 kg

Data analytics mcq with answers

20. What percentage of measurements in a dataset fall above the median?

- 1. 49%
- 2. 50%
- 3. 51%
- 4. Cannot Be Determined

Show Answer

Cannot Be Determined

21. The proportion of variation in 5k race times that can be explained by the variation in the age of competitive male runners was approximately 0.663. What is the value of the sample linear correlation coefficient? Round to 3 decimal places.

- 1. 0.663
- 2. 0.814
- 3. -0.814
- 4. 0.440

Show Answer

-0.814

22. Using all of the results provided, is it reasonable to predict the 5k race time (minutes) of a competitive male runner 73 years of age?"

Data Analytics Mcq with Answers

1. Yes; linear correlation between age and 5k race times is statistically significant
2. Yes; both the sample linear regression equation and an age in years is provided
3. No; linear correlation between age and 5k race times is not statistically significant
4. No; the age provided is beyond the scope of our available sample data” d

Show Answer

No; linear correlation between age and 5k race times is not statistically significant

23. 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 and 2

Show Answer

Both 1 and 2

24. Algorithm is

1. It uses machine-learning techniques. Here program can learn from past experience and adapt themselves to new situations
2. Computational procedure that takes some value as input and produces some value as output
3. Science of making machines performs tasks that would require intelligence when performed by humans
4. None of these

Show Answer

Data Analytics Mcq with Answers

Computational procedure that takes some value as input and produces some value as output

25. Bias 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

Show Answer

Any mechanism employed by a learning system to constrain the search space of a hypothesis

26. Classification is

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

Show Answer

A subdivision of a set of examples into a number of classes

27. Binary attribute are

1. This takes only two values. In general, these values will be 0 and 1 and .they can be coded as one bit
2. The natural environment of a certain species
3. Systems that can be used without knowledge of internal operations

Data Analytics Mcq with Answers

4. None of these

Show Answer

This takes only two values. In general, these values will be 0 and 1 and .they can be coded as one bit

28. Cluster is

1. Group of similar objects that differ significantly from other objects
2. Operations on a database to transform or simplify data in order to prepare it for a machine-learning algorithm
3. Symbolic representation of facts or ideas from which information can potentially be extracted
4. None of these

Show Answer

Group of similar objects that differ significantly from other objects

29. A definition of a concept is _____ if it recognizes all the instances of that concept

1. Complete
2. Consistent
3. Constant
4. None of these

Show Answer

Complete

30. A definition of a concept is _____ if it classifies any examples as coming within the concept

1. Complete
2. Consistent
3. Constant

Data Analytics Mcq with Answers

4. None of these

Show Answer

Consistent

30. Data selection is

1. The actual discovery phase of a knowledge discovery process
2. The stage of selecting the right data for a KDD process
3. A subject-oriented integrated time variant non-volatile collection of data in support of management
4. None of these

Show Answer

The stage of selecting the right data for a KDD process

30. 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

Show Answer

The task of assigning a classification to a set of examples

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

1. industry statistics
2. economic statistics
3. applied statistics

Data Analytics Mcq with Answers

4. applied statistics

Show Answer

applied statistics

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

1. answering yes/no questions about the data
2. estimating numerical characteristics of the data
3. modeling relationships within the data
4. describing associations within the data

Show Answer

modeling relationships within the data

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

1. True
2. False
3. Can be true or False
4. Can not say

Show Answer

TRUE

4. What is a hypothesis?

1. A statement that the researcher wants to test through the data collected in a study.
2. A research question the results will answer.
3. A theory that underpins the study.
4. A statistical method for calculating the extent to which the results could have happened by chance.

Show Answer

Data Analytics Mcq with Answers

A statement that the researcher wants to test through the data collected in a study.

5. What is the cyclical process of collecting and analysing data during a single research study called?

1. Interim Analysis
2. Inter analysis
3. inter item analysis
4. constant analysis

Show Answer

Interim Analysis

6. The process of quantifying data is referred to as _____

1. Topology
2. Diagramming
3. Enumeration
4. coding

Show Answer

Enumeration

7. An advantage of using computer programs for qualitative data is that they _

1. Can reduce time required to analyse data (i.e., after the data are transcribed)
2. Help in storing and organising data
3. Make many procedures available that are rarely done by hand due to time constraints
4. All of the above

Show Answer

Data Analytics Mcq with Answers

All of the Above

8. Boolean operators are words that are used to create logical combinations.

1. True
2. False

Show Answer

True

9. _____ are the basic building blocks of qualitative data.

1. Categories
2. Units
3. Individuals
4. None of the above

Show Answer

Categories

10. This is the process of transforming qualitative research data from written interviews or field notes into typed text.

1. Segmenting
2. Coding
3. Transcription
4. Mnemoning

Show Answer

Transcription

data analytics mcq with answers pdf

11. A challenge of qualitative data analysis is that it often includes data that are unwieldy and complex; it is a major challenge to make sense of the large pool of data.

Data Analytics Mcq with Answers

1. True
2. False

Show Answer

True

12. Hypothesis testing and estimation are both types of descriptive statistics.

1. True
2. False

Show Answer

False

13. A set of data organised in a participants(rows)-by-variables(columns) format is known as a “data set.”

1. True
2. False

Show Answer

True

14. A graph that uses vertical bars to represent data is called a ____

1. Line graph
2. Bar graph
3. Scatterplot
4. Vertical graph

Show Answer

Bar graph

15. ____ are used when you want to visually examine the relationship between two quantitative variables.

Data Analytics Mcq with Answers

1. Bar graph
2. pie graph
3. line graph
4. Scatterplot

Show Answer

Scatterplot

16. The denominator (bottom) of the z-score formula is

1. The standard deviation
2. The difference between a score and the mean
3. The range
4. The mean

Show Answer

The standard deviation

17. Which of these distributions is used for a testing hypothesis?

1. Normal Distribution
2. Chi-Squared Distribution
3. Gamma Distribution
4. Poisson Distribution

Show Answer

Chi-Squared Distribution

18. A statement made about a population for testing purpose is called?

1. Statistic
2. Hypothesis

Data Analytics Mcq with Answers

3. Level of Significance
4. Test-Statistic

Show Answer

Hypothesis

19. If the assumed hypothesis is tested for rejection considering it to be true is called?

1. Null Hypothesis
2. Statistical Hypothesis
3. Simple Hypothesis
4. Composite Hypothesis

Show Answer

Null Hypothesis

20. If the null hypothesis is false then which of the following is accepted?

1. Null Hypothesis
2. Positive Hypothesis
3. Negative Hypothesis
4. Alternative Hypothesis.

Show Answer

Alternative Hypothesis.

21. Alternative Hypothesis is also called as?

1. Composite hypothesis
2. Research Hypothesis
3. Simple Hypothesis
4. Null Hypothesis

Show Answer

Data Analytics Mcq with Answers

Research Hypothesis

1. What is the minimum no. of variables/ features required to perform clustering?

- 1. 0
- 2. 1
- 3. 2
- 4. 3

Show Answer

1

2. For two runs of K-Mean clustering is it expected to get same clustering results?

- 1. Yes
- 2. No

Show Answer

No

3. 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

Show Answer

K-means clustering algorithm

4. The discrete variables and continuous variables are two types of

- 1. Open end classification

Data Analytics Mcq with Answers

2. Time series classification
3. Qualitative classification
4. Quantitative classification

Show Answer

Quantitative classification

5. 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

Show Answer

A class of learning algorithm that tries to find an optimum classification of a set of examples using the probabilistic theory.

6. 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

Show Answer

Measure of the accuracy, of the classification of a concept that is given by a certain theory

Data Analytics Mcq with Answers

7. 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 above

Show Answer

The distance between two points as calculated using the Pythagoras theorem

8. Hybrid is

1. Combining different types of method or information
2. Approach to the design of learning algorithms that is structured along the lines of the theory of evolution.
3. Decision support systems that contain an information base filled with the knowledge of an expert formulated in terms of if-then rules.
4. none of above

Show Answer

Combining different types of method or information

9. 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 algorithm

Show Answer

Data Analytics Mcq with Answers

Greedy Algorithms

10. Discovery is

1. It is hidden within a database and can only be recovered if one is given certain clues (an example IS encrypted information).
2. The process of executing implicit previously unknown and potentially useful information from data
3. An extremely complex molecule that occurs in human chromosomes and that carries genetic information in the form of genes.
4. None of these

Show Answer

The process of executing implicit previously unknown and potentially useful information from data

11. Hidden knowledge referred to

1. A set of databases from different vendors, possibly using different database paradigms
2. An approach to a problem that is not guaranteed to work but performs well in most cases
3. Information that is hidden in a database and that cannot be recovered by a simple SQL query.
4. None of these

Show Answer

Information that is hidden in a database and that cannot be recovered by a simple SQL query.

12. Decision trees cannot handle categorical attributes with many distinct values, such as country codes for telephone numbers.

1. True
2. False

Show Answer

Data Analytics Mcq with Answers

False

13. Enrichment 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

Show Answer

A stage of the KDD process in which new data is added to the existing selection

14. _____ 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 above

Show Answer

Naïve Bayes classifiers

15. High entropy means that the partitions in classification are

1. Pure
2. Not Pure
3. Usefull
4. useless

Show Answer

Data Analytics Mcq with Answers

Uses a single processor or computer

16. 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

Show Answer

Attributes are statistically dependent of one another given the class value.

17. 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

Show Answer

Max Entropy is 3

18. 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

Show Answer

k-nearest neighbor is same as k-means

Data Analytics Mcq with Answers

19. 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
3. Regression
4. None of these

Show Answer

Clustering

20. Can we use K Mean Clustering to identify the objects in video?

1. Yes
2. No

Show Answer

Yes

21. 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

Show Answer

Unsupervised

22. _____ metric is examined to determine a reasonably optimal value of k.

1. Mean Square Error
2. Within Sum of Squares (WSS)

Data Analytics Mcq with Answers

- 3. Speed
- 4. None of these

Show Answer

Within Sum of Squares (WSS)

23. 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 and 2

Show Answer

Both 1 and 2

24. 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

Show Answer

1

25. Confidence is a measure of how X and Y are really related rather than coincidentally happening together.

- 1. True
- 2. False

Show Answer

Data Analytics Mcq with Answers

False

26. _____ 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

Show Answer

Collaborative Filtering Systems

27. There are _____ major Classification of Collaborative Filtering Mechanisms

1. 1
2. 2
3. 3
4. none of above

Show Answer

2

28. Movie Recommendation to people is an example of

1. User Based Recommendation
2. Item Based Recommendation
3. Knowledge Based Recommendation
4. content based recommendation

Show Answer

Item Based Recommendation

Data Analytics Mcq with Answers

29. _____ recommenders rely on an explicitly defined set of recommendation rules

1. Constraint Based
2. Case Based
3. Content Based
4. User Based

Show Answer

Case Based

30. Parallelized hybrid recommender systems operate independently of one another and produce separate recommendation lists.

1. True
2. False

Show Answer

False

1. Precies and steady format data is_____

1. Structured Data
2. UnStructured Data
3. semi Structured Data
4. Quasi Structured Data

Show Answer

Structured Data

2. Inconsistant Data is_____

1. Structured Data
2. Un Structured Data
3. semi Structured Data

Data Analytics Mcq with Answers

4. Quasi Structured Data

Show Answer

Un Structured Data

3. Format that self defines itself is_____

1. Structured Data
2. Un Structured Data
3. semi Structured Data
4. Quasi Structured Data

Show Answer

semi Structured Data

4. A little Bit inconsistant data is_____

1. Structured Data
2. Un Structured Data
3. semi Structured Data
4. Quasi Structured Data

Show Answer

Quasi Structured Data

5. XML is an example of_____

1. Structured Data
2. UnStructured Data
3. semi Structured Data
4. Quasi Structured Data

Show Answer

semi Structured Data

Data Analytics Mcq with Answers

6. RDBMS Follows_____

1. Structured Data
2. Un Structured Data
3. semi Structured Data
4. Quasi Structured Data

Show Answer

Structured Data

7. Watson is developed by_____

1. IBM
2. Microsoft
3. AT&T
4. Google

Show Answer

IBM

8. Hadoop is _ based Framework.

1. C++
2. Python
3. JAVA
4. C#

Show Answer

JAVA

9. Which of the following are components of Hadoop?

1. MAPREDUCE
2. YARN

Data Analytics Mcq with Answers

- 3. HDFS
- 4. All of Above

Show Answer

All of Above

10. Which of the following are components of HIVE?

- 1. JDBC
- 2. Thrift Server
- 3. CLI
- 4. All of Above

Show Answer

All of Above

11. Mahout provides_____

- 1. JAVA Executable Libraries
- 2. C# Executables
- 3. Mountable Image Format
- 4. All of Above

Show Answer

JAVA Executable Libraries

12. Which of the following are components of HIVE?

- 1. FLATTEN
- 2. Thrift Server
- 3. Muster
- 4. None of these

Show Answer

Data Analytics Mcq with Answers

Thrift Server

13. Which of the following is components of Hadoop?

1. Fork
2. YARN
3. CLI
4. Metadata

Show Answer

YARN

14. Which of the following is a clustering techique?

1. Fuzzy K means
2. Canopy
3. K-Means
4. All of above

Show Answer

all of above

15. Which of the following is HBASE Data Model Terminology?

1. Row
2. Table
3. Column
4. All of Above

Show Answer

all of above

16. Which of the following is not a classification techique?

1. Logistic Regression

Data Analytics Mcq with Answers

2. Random Forest
3. Recommender Algo
4. Naïve Bayes

Show Answer

Recommender Algo

17. Which of the following is a classification technique?

1. Logistic Regression
2. Random Forest
3. Naïve Bayes
4. All of Above

Show Answer

all of above

18. Which of the following is HBASE Data Model Terminology?

1. Column Family
2. Cell
3. Timestamp
4. All of Above

Show Answer

All of above

19. Which of the following is a clustering technique?

1. Logistic Regression
2. Random Forest
3. K-Means
4. Naïve Bayes

Data Analytics Mcq with Answers

Show Answer

K-Means

20. Which of the following is HBASE Data Model Terminology?

1. Identifier
2. Variant
3. Timestamp
4. None of the above

Show Answer

Timestamp

21. Which of the following is not a classification technique?

1. Logistic Regression
2. Random Forest
3. K-Means
4. Naïve Baye

Show Answer

K-Means

22. Which of the following is HBASE Data Model Terminology?

1. Identifier
2. Variant
3. Column Qualifier
4. None of the above

Show Answer

Column Qualifier

23. Which of the following is not a clustering techique?

Data Analytics Mcq with Answers

1. Logistic Regression
2. Canopy
3. K-Means
4. Fuzzy K means

Show Answer

Logistic Regression

24. Point out the correct statement.

1. Hadoop do need specialized hardware to process the data
2. Hadoop 2.0 allows live stream processing of real-time data
3. In Hadoop programming framework output files are divided into lines or records
4. None of the above

Show Answer

Hadoop 2.0 allows live stream processing of real-time data

25. What was Hadoop named after?

1. Creator Doug Cutting's favorite circus act
2. Cutting's high school rock band
3. The toy elephant of Cutting's son
4. A sound Cutting's laptop made during Hadoop development

Show Answer

The toy elephant of Cutting's son

26. _____ programming model used to develop Hadoop-based applications that can process massive amounts of data.

1. MapReduce
2. Mahout

Data Analytics Mcq with Answers

- 3. Oozie
- 4. None of the above

Show Answer

MapReduce

27. Hadoop is a framework that works with a variety of related tools. Common cohorts include __

- 1. MapReduce, Hive and HBase
- 2. MapReduce, MySQL and Google Apps
- 3. MapReduce, Hummer and Iguana
- 4. All of above

Show Answer

MapReduce, Hive and HBase

28. NoSQL databases is used mainly for handling large volumes of __ data.

- 1. Structured Data
- 2. Un Structured Data
- 3. semi Structured Data
- 4. Quasi Structured Data

Show Answer

Un Structured Data

29. Which of the following is not a phase of Data Analytics Life Cycle?

- 1. Communication
- 2. Recall
- 3. Data Preparation
- 4. Model Planning

Show Answer

Data Analytics Mcq with Answers

Recall

30. Which of the following is a NoSQL Database Type?

1. SQL
2. Document databases
3. JSON
4. All of above

Show Answer

Document databases

30. Which of the following is not a NoSQL database

1. SQL Server
2. MongoDB
3. Cassandra
4. None of the above

Show Answer

SQL Server

1. Data visualization is related with

1. Pictorial representations
2. numerical representation
3. numerical calculations
4. None of these

Show Answer

Pictorial representations

2. Which of the following are Use of data visualization

Data Analytics Mcq with Answers

1. See context of data
2. Clear data understanding
3. finding pattern in data
4. all of above

Show Answer

all of above

3. Which of the following statements are true about using visualizations to display a dataset?

- I. Visualizations are visually appealing, but don't help the viewer understand relationships that exist in the data
- II. Visualizations like graphs, charts, or visualizations with pictures are useful for conveying information, while tables just filled with text are not useful.
- III. Patterns that exist in the data can be found more easily by using a visualization

1. I AND II
2. II AND III
3. I AND III
4. ONLY III

Show Answer

ONLY III

4. The plot method on Series and DataFrame is just a simple wrapper around __

1. gplt.plot()
2. plt.plot()
3. plt.plotgraph()
4. none of the mentioned

Data Analytics Mcq with Answers

Show Answer

plt.plot()

5. Point out the correct combination with regards to kind keyword for graph plotting.

1. ‘hist’ for histogram
2. ‘box’ for boxplot
3. ‘area’ for area plots
4. all of the mentioned

Show Answer

all of the mentioned

6. Which of the following value is provided by kind keyword for barplot?

1. bar
2. kde
3. hexbin
4. none of the mentioned

Show Answer

bar

7. You can create a scatter plot matrix using the __ method in pandas.tools.plotting.

1. sca_matrix
2. scatter_matrix
3. DataFrame.plot
4. all of the mentioned

Show Answer

scatter_matrix

Data Analytics Mcq with Answers

8. Plots may also be adorned with error bars or tables.

1. True
2. FALSE

Show Answer

True

9. Which of the following plots are often used for checking randomness in time series?

1. Autocausation
2. Autorank
3. Autocorrelation
4. none of the mentioned

Show Answer

Autocorrelation

10. __ plots are used to visually assess the uncertainty of a statistic

1. Lag
2. RadViz
3. Bootstrap
4. All Above

Show Answer

Bootstrap

Data analytics mcq with answers

11. Which of the following is not a challenge in Big Data Visualization?

1. Velocity
2. Volume
3. Version

Data Analytics Mcq with Answers

4. Variety

Show Answer

Version

12. Which of the following is not a problem in Big Data Visualization

1. Visual Noise
2. Scaled Data
3. Large image perception
4. Information Loss

Show Answer

Scaled Data

13. Which of the following is a problem in Big Data Visualization

1. Structured Data
2. Scaled Data
3. Visual Noise
4. Multiple valued Data

Show Answer

Visual Noise

14. Which of the candidate is suitable for interactive visualtization?

1. Type of Visual
2. Cardinality
3. Size of data
4. all of above

Show Answer

all of above

Data Analytics Mcq with Answers

15. Which of the following follows interactive visualization approach?

1. Zoom+Pan
2. Focus+Context
3. Overview+Details
4. all of above

Show Answer

all of above

16. Visual Mapping is important for_____

1. Remapping
2. Overview+Details
3. Focus
4. Context

Show Answer

Remapping

17. Data visualtization techniques are:

1. Scatter Plot
2. Line Chart
3. Pie Chart
4. all of above

Show Answer

all of above

18. Information Visualtization techniques are

1. Flow Chart
2. Time Line

Data Analytics Mcq with Answers

- 3. DFD
- 4. All of above

Show Answer

All of above

19. Which of the following is related term with correlation?

- 1. Exponential
- 2. U-Shape
- 3. Null
- 4. All of above

Show Answer

All of above

20. Column graph is another name for _

- 1. Bar Chart
- 2. Scatterplot
- 3. Histogram
- 4. Area Chart

Show Answer

Bar Chart

21. Which of the following is category of timeline?

- 1. Linear Timeline
- 2. Modular Timeline
- 3. Variant Timeline
- 4. ER Timeline

Show Answer

Data Analytics Mcq with Answers

Linear Timeline

22. Which of the following specifies relationship amongst variables?

1. Scatter Plot
2. Line Chart
3. Area Chart
4. All of above

Show Answer

All of above

23.Which of the following specifies category Proportions?

1. Pie Chart
2. Histogram
3. Bar chart
4. All of above

Show Answer

All of above

24. Which of the following is category of timeline?

1. Variant Timeline
2. ER Timeline
3. Comarative Timeline
4. Modular Timeline

Show Answer

Comarative Timeline

25. Information Visualtization techniques are

1. Semantic Network

Data Analytics Mcq with Answers

2. Histogram
3. Area Chart
4. None of these

Show Answer

Semantic Network

26. Information Visualtization techniques are

1. Scatter Plot
2. Time Line
3. Bubble Chart
4. None of these

Show Answer

Time Line

27. Information Visualtization techniques are

1. Flow Chart
2. Line Chart
3. Pie Chart
4. None of these

Show Answer

Flow Chart

28. Which of the following are Use of data visualtization

1. See context of data
2. Clear data understanding
3. finding pattern in data
4. all of above

Data Analytics Mcq with Answers

Show Answer

all of above

29. Which of the following specifies relationship amongst variables?

1. Pie Chart
2. Histogram
3. Area Chart
4. None of these

Show Answer

Area Chart

30. Which of the following specifies category Proportions?

1. Pie Chart
2. Scatter Plot
3. Line Chart
4. None of these

Show Answer

Pie Chart

1. 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.

1. Decision tree
2. Graphs
3. Trees
4. Neural Networks

Show Answer

Decision tree

Data Analytics Mcq with Answers

2. What is Decision Tree?

1. Flow-Chart
2. Structure in which internal node represents test on an attribute, each branch represents outcome of test and each leaf node represents class label
3. 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
4. None of Above

Show Answer

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

3. Decision Trees can be used for Classification Tasks.

1. TRUE
2. FALSE

Show Answer

TRUE

4. Choose from the following that are Decision Tree nodes?

1. Decision Nodes
2. End Nodes
3. Chance Nodes
4. All of Above

Show Answer

All of Above

5. Decision Nodes are represented by __

1. Disks

Data Analytics Mcq with Answers

2. Squares
3. Circles
4. Triangles

Show Answer

Squares

6. Chance Nodes are represented by __

1. Disks
2. Squares
3. Circles
4. Triangles

Show Answer

Circles

7. End Nodes are represented by __

1. Disks
2. Squares
3. Circles
4. Triangles

Show Answer

Triangles

8. Which of the following are the advantage/s of Decision Trees?

1. Possible Scenarios can be added
2. Use a white box model, If given result is provided by a model
3. Worst, best and expected values can be determined for different scenarios
4. All of Above

Data Analytics Mcq with Answers

Show Answer

All of Above

9. 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

Show Answer

Attributes are statistically dependent of one another given the class value.

10. Which of the following is not supervised learning?

1. Clustering
2. Decision Tree
3. Linear Regression
4. Naive Bayesian

Show Answer

Clustering

Data analytics mcq with answers

11. How many terms are required for building a bayes model?

1. 1
2. 2
3. 3
4. 4

Show Answer

Data Analytics Mcq with Answers

12. Where does the bayes rule can be used?

1. Solving queries
2. Increasing complexity
3. Decreasing complexity
4. Answering probabilistic query

Show Answer

Answering probabilistic query

13. How the bayesian network can be used to answer any query?

1. Full distribution
2. Joint distribution
3. Partial distribution
4. All of Above

Show Answer

Joint distribution

14. What is the consequence between a node and its predecessors while creating bayesian network?

1. Functionally dependent
2. Dependant
3. Conditionally independent
4. Both Conditionally dependant & Dependant

Show Answer

Conditionally independent

15. 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.

Data Analytics Mcq with Answers

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

Show Answer

A class of learning algorithm that tries to find an optimum classification of a set of examples using the probabilistic theory.

16. Bias 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

Show Answer

Any mechanism employed by a learning system to constrain the search space of a hypothesis

17. Background knowledge referred to

1. Additional acquaintance used by a learning algorithm to facilitate the learning process
2. A neural network that makes use of a hidden layer
3. It is a form of automatic learning.
4. None of these

Data Analytics Mcq with Answers

Show Answer

Additional acquaintance used by a learning algorithm to facilitate the learning process

18. Discriminating between spam and ham e-mails is a classification task

1. TRUE
2. FALSE

Show Answer

TRUE

19. which of the following is not involve in data mining?

1. Knowledge extraction
2. Data archaeology
3. Data exploration
4. Data transformation

Show Answer

Data transformation

20. Naive prediction is

1. A class of learning algorithms that try to derive a Prolog program from examples
2. A table with n independent attributes can be seen as an n- dimensional space.
3. A prediction made using an extremely simple method, such as always predicting the same output.
4. None of these

Show Answer

A prediction made using an extremely simple method, such as always predicting the same output.

Data Analytics Mcq with Answers

21. Node is _____

1. A component of a network
2. In the context of KDD and data mining, this refers to random errors in a database table.
3. One of the defining aspects of a data warehouse
4. None of these

Show Answer

A component of a network

22. Prediction is

1. The result of the application of a theory or a rule in a specific case
2. One of several possible entries within a database table that is chosen by the designer as the primary means of accessing the data in the table.
3. Discipline in statistics that studies ways to find the most interesting projections of multi-dimensional spaces.
4. None of these

Show Answer

The result of the application of a theory or a rule in a specific case

23. What is the relation between the distance between clusters and the corresponding class discriminability?

1. proportional
2. inversely-proportional
3. no-relation
4. None of these

Show Answer

proportional

Data Analytics Mcq with Answers

24. the classification method in which the upper limit of interval is same as of lower class interval is called

1. exclusive method
2. inclusive method
3. mid point method
4. None of these

Show Answer

exclusive method

25. larger value is 60 and the smallest value is 40 and the number of classes is 5 then the class interval is

1. 20
2. 25
3. 4
4. 15

Show Answer

4

26. summary and presentation of data in tabular form with several non overlapping classes is referred as

1. nominal distribution
2. frequency distribution
3. ordinal distribution
4. None of these

Show Answer

frequency distribution

27. the classification method in which the upper and lower limit of interval is also in class interval itself is called

Data Analytics Mcq with Answers

1. exclusive method
2. inclusive method
3. mid point method
4. None of these

Show Answer

inclusive method

28. Suppose there are 25 base classifiers. Each classifier has error rates of $e = 0.35$. Suppose you are using averaging as ensemble of above 25 classifiers will make a wrong prediction? Note: all classifiers are independent of each other

1. 0.05
2. 0.06
3. 0.07
4. 0.08

Show Answer

0.06

29. The most widely used metrics and tools to assess a classification model are:

1. Confusion matrix
2. Cost-sensitive accuracy
3. Area under the ROC curve
4. All of Above

Show Answer

All of Above

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

Data Analytics Mcq with Answers

1. Normalize the data → PCA → training
2. PCA → normalize PCA output → training
3. Normalize the data → PCA → normalize PCA output → training
4. None of these

Show Answer

Normalize the data → PCA → training

31. Which of the following is true about Naive Bayes ?

1. Assumes that all the features in a dataset are equally important
2. Assumes that all the features in a dataset are independent
3. both a and b
4. None of these

Show Answer

both a and b

32. 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

1. 1 and 2
2. 2 and 3
3. 1, 2, and 3
4. 1 and 3

Show Answer

1, 2, and 3

Data Analytics Mcq with Answers

1. 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. 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.

1. K-Means clustering
2. conceptual clustering
3. expectation maximization
4. agglomerative clustering

Show Answer

K-Means clustering

2. The correlation coefficient for two real-valued attributes is -0.85. What does this value tell you?

1. The attributes are not linearly related.
2. As the value of one attribute decreases the value of the second attribute increases.
3. As the value of one attribute increases the value of the second attribute also increases.
4. The attributes show a linear relationship

Show Answer

As the value of one attribute decreases the value of the second attribute increases.

3. Given a rule of the form IF X THEN Y, rule confidence is defined as the conditional probability that

1. Y is false when X is known to be false.
2. Y is true when X is known to be true.
3. X is true when Y is known to be true

Data Analytics Mcq with Answers

4. X is false when Y is known to be false.

Show Answer

Y is true when X is known to be true.

4. Chameleon is

1. Density based clustering algorithm
2. Partitioning based algorithm
3. Model based algorithm
4. Hierarchical clustering algorithm

Show Answer

Hierarchical clustering algorithm

5. Find odd man out

1. DBSCAN
2. K-Mean
3. PAM
4. None of above

Show Answer

DBSCAN

6. The number of iterations in apriori _

1. increases with the size of the data
2. decreases with the increase in size of the data
3. increases with the size of the maximum frequent set
4. decreases with increase in size of the maximum frequent set

Show Answer

increases with the size of the maximum frequent set

Data Analytics Mcq with Answers

7. Which of the following are interestingness measures for association rules?

1. Recall ‘
2. Lift
3. Accuracy
4. All of Above

Show Answer

Lift

8. Given a frequent itemset L, If $|L| = k$, then there are

1. $2k - 1$ candidate association rules
2. $2k$ candidate association rules
3. $2k - 2$ candidate association rules
4. $2k - 2$ candidate association rules

Show Answer

$2k - 2$ candidate association rules (2 to power $k - 2$)

9. _____ is an example for case based-learning

1. Decision trees
2. Neural networks
3. Genetic algorithm
4. K-nearest neighbor

Show Answer

K-nearest neighbor

10. The average positive difference between computed and desired outcome values.

1. mean positive

Data Analytics Mcq with Answers

2. error mean squared
3. error mean absolute
4. error root mean squared error

Show Answer

error mean absolute

11. Frequent item sets is

1. Superset of only closed frequent item sets
2. Superset of only maximal frequent item sets
3. Subset of maximal frequent item sets
4. Superset of both closed frequent item sets and maximal frequent item sets

Show Answer

Superset of both closed frequent item sets and maximal frequent item sets

12. Assume that we have a dataset containing information about 200 individuals. A supervised data mining session has discovered the following rule: IF age < 30 & credit card insurance = yes THEN life insurance = yes Rule Accuracy: 70% and Rule Coverage: 63% How many individuals in the class life insurance= no have credit card insurance and are less than 30 years old?

1. 63
2. 38
3. 40
4. 89

Show Answer

38

13. Which of the following is cluster analysis?

1. Simple segmentation

Data Analytics Mcq with Answers

2. Grouping similar objects
3. Labeled classification
4. Query results grouping

Show Answer

Grouping similar objects

14. A good clustering method will produce high quality clusters with

1. high inter class similarity
2. high intra class similarity
3. low intra class similarity
4. None of above

Show Answer

low intra class similarity

15. Which two parameters are needed for DBSCAN

1. Min threshold
2. Min points and eps
3. Min sup and min confidence
4. Number of centroids

Show Answer

Min points and eps

16. Which statement is true about neural network and linear regression models?

1. Both techniques build models whose output is determined by a linear sum of weighted input attribute values.
2. The output of both models is a categorical attribute value.
3. Both models require numeric attributes to range between 0 and 1.

Data Analytics Mcq with Answers

4. Both models require input attributes to be numeric.

Show Answer

Both models require input attributes to be numeric.

17. In Apriori algorithm, if 1 item-sets are 100, then the number of candidate 2 item-sets are

1. 100
2. 200
3. 4950
4. 5000

Show Answer

4950

18. Significant Bottleneck in the Apriori algorithm is

1. Finding frequent itemsets
2. Pruning
3. Candidate generation
4. Number of iterations

Show Answer

Candidate generation

19. Machine learning techniques differ from statistical techniques in that machine learning methods

1. are better able to deal with missing and noisy data
2. typically assume an underlying distribution for the data
3. have trouble with large-sized datasets
4. are not able to explain their behavior.

Show Answer

Data Analytics Mcq with Answers

are better able to deal with missing and noisy data

20. The probability of a hypothesis before the presentation of evidence.

1. a priori
2. posterior
3. conditional
4. subjective

Show Answer

a priori

21. KDD represents extraction of

1. data
2. knowledge
3. rules
4. model

Show Answer

knowledge

21. Which statement about outliers is true?

1. Outliers should be part of the training dataset but should not be present in the test data.
2. Outliers should be identified and removed from a dataset.
3. The nature of the problem determines how outliers are used
4. Outliers should be part of the test dataset but should not be present in the training data.

Show Answer

The nature of the problem determines how outliers are used

21. The most general form of distance is

Data Analytics Mcq with Answers

1. Manhattan
2. Euclidian
3. Mean
4. Minkowski

Show Answer

Minkowski

21. Which Association Rule would you prefer

1. High support and medium confidence
2. High support and low confidence
3. Low support and high confidence
4. Low support and low confidence

Show Answer

Low support and high confidence

21. In a Rule based classifier, If there is a rule for each combination of attribute values, what do you called that rule set R

1. Exhaustive
2. Inclusive
3. Comprehensive
4. Mutually exclusive

Show Answer

Exhaustive

21. The apriori property means

1. If a set cannot pass a test, its supersets will also fail the same test
2. To decrease the efficiency, do level-wise generation of frequent item sets
3. To improve the efficiency, do level-wise generation of frequent item sets

Data Analytics Mcq with Answers

4. If a set can pass a test, its supersets will fail the same test

Show Answer

If a set cannot pass a test, its supersets will also fail the same test

21. If an item set ‘XYZ’ is a frequent item set, then all subsets of that frequent item set are

1. Undefined
2. Not frequent
3. Frequent
4. Can not say

Show Answer

Frequent

21. The probability that a person owns a sports car given that they subscribe to automotive magazine is 40%. We also know that 3% of the adult population subscribes to automotive magazine. The probability of a person owning a sports car given that they don't subscribe to automotive magazine is 30%. Use this information to compute the probability that a person subscribes to automotive magazine given that they own a sports car

1. 0.0368
2. 0.0396
3. 0.0389
4. 0.0398

Show Answer

0.0396

21. Simple regression assumes a __ relationship between the input attribute and output attribute.

1. quadratic
2. inverse

Data Analytics Mcq with Answers

- 3. linear
- 4. reciprocal

Show Answer

linear

21. To determine association rules from frequent item sets

- 1. Only minimum confidence needed
- 2. Neither support nor confidence needed
- 3. Both minimum support and confidence are needed
- 4. Minimum support is needed

Show Answer

Both minimum support and confidence are needed

21. If {A,B,C,D} is a frequent itemset, candidate rules which is not possible is

- 1. C → A
- 2. D → ABCD
- 3. A → BC
- 4. B → ADC

Show Answer

D → ABCD

21. Classification rules are extracted from _

- 1. decision tree
- 2. root node
- 3. branches
- 4. siblings

Show Answer

Data Analytics Mcq with Answers

decision tree

21. What does K refers in the K-Means algorithm which is a non-hierarchical clustering approach?

1. Complexity
2. Fixed value
3. No of iterations
4. number of clusters

Show Answer

number of clusters

21. If Linear regression model perfectly fit i.e., train error is zero, then _____

1. Test error is also always zero
2. Test error is non zero
3. Couldn't comment on Test error
4. Test error is equal to Train error

Show Answer

Couldn't comment on Test error

21. How many coefficients do you need to estimate in a simple linear regression model (One independent variable)?

1. 1
2. 2
3. 3
4. 4

Show Answer

2

Data Analytics Mcq with Answers

21. In a simple linear regression model (One independent variable), If we change the input variable by 1 unit. How much output variable will change?

1. by 1
2. no change
3. by intercept
4. by its slope

Show Answer

by its slope

21. In syntax of linear model lm(formula,data,..), data refers to __

1. Matrix
2. array
3. vector
4. list

Show Answer

vector

21. In the mathematical Equation of Linear Regression $Y = \beta_1 + \beta_2X + \epsilon$, (β_1, β_2) refers to __

1. (X-intercept, Slope)
2. (Slope, X-Intercept)
3. (Y-Intercept, Slope)
4. (slope, Y-Intercept)

Show Answer

(Y-Intercept, Slope)

Data Analytics Mcq with Answers

1. Which of the following is not an example of Social Media?

1. Twitter
2. Google
3. Instagram
4. Youtube

Show Answer

Google

2. By 2025, the volume of digital data will increase to

1. TB
2. YB
3. ZB
4. EB

Show Answer

ZB

3. Data Analysis is a process of

1. inspecting data
2. cleaning data
3. transforming data
4. All of Above

Show Answer

All of above

4. Does Facebook uses “Big Data ” to perform the concept of Flashback?

1. True
2. False

Data Analytics Mcq with Answers

Show Answer

True

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

1. Data Mining
2. Predictive Intelligence
3. Business Intelligence
4. Text Analytics

Show Answer

Predictive Intelligence

6. 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

Show Answer

Big data

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

1. 2
2. 3
3. 4
4. 5

Show Answer

2

Data Analytics Mcq with Answers

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

1. Integer descriptor
2. floating descriptor
3. numerical descriptor
4. decimal descriptor

Show Answer

numerical descriptor

9. ____ have a structure but cannot be stored in a database.

1. Structured
2. Semi Structured
3. Unstructured
4. None of these

Show Answer

None of these

10. Data generated from online transactions is one of the example for volume of big data

1. TRUE
2. FALSE

Show Answer

TRUE

11. Velocity is the speed at which the data is processed

1. True
2. False

Show Answer

Data Analytics Mcq with Answers

False

12. Value tells the trustworthiness of data in terms of quality and accuracy

1. TRUE
2. FALSE

Show Answer

False

13. Hortonworks was introduced by Cloudera and owned by Yahoo

1. True
2. False

Show Answer

False

14. ____ refers to the ability to turn your data useful for business

1. Velocity
2. variety
3. Value
4. Volume

Show Answer

Value

15. GFS consists of ____ Master and ____ Chunk Servers

1. Single, Single
2. Multiple, Single
3. Single, Multiple
4. Multiple, Multiple

Show Answer

Data Analytics Mcq with Answers

Single, Multiple

16. Data Analysis is defined by the statistician?

1. William S.
2. Hans Peter Luhn
3. Gregory Piatetsky-Shapiro
4. John Tukey

Show Answer

John Tukey

17. Files are divided into ____ sized Chunks.

1. Static
2. Dynamic
3. Fixed
4. Variable

Show Answer

Fixed

18. ____ is an open source framework for storing data and running application on clusters of commodity hardware.

1. HDFS
2. Hadoop
3. MapReduce
4. Cloud

Show Answer

Hadoop

19. HDFS Stores how much data in each clusters that can be scaled at any time?

Data Analytics Mcq with Answers

1. 32
2. 64
3. 128
4. 256

Show Answer

128

20. Hadoop MapReduce allows you to perform distributed parallel processing on large volumes of data quickly and efficiently... is this MapReduce or Hadoop

1. True
2. False

Show Answer

True

data analytics mcq questions and answers

21. Google Introduced MapReduce Programming model in 2004

1. True
2. False

Show Answer

True

22. Hadoop YARN is used for Cluster Resource Management in Hadoop Ecosystem

1. True
2. False

Show Answer

True

Data Analytics Mcq with Answers

23. _____ phase sorts the data & _____ creates logical clusters.

1. Reduce, YARN
2. MAP, YARN
3. REDUCE, MAP
4. MAP, REDUCE

Show Answer

MAP, REDUCE

24. There is only one operation between Mapping and Reducing

1. True
2. False

Show Answer

True

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

1. answering yes/no questions about the data
2. estimating numerical characteristics of the data
3. describing associations within the data
4. modeling relationships within the data

Show Answer

answering yes/no questions about the data

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

1. True
2. False

Show Answer

True

Data Analytics Mcq with Answers

27. _____ is factors considered before Adopting Big Data Technology

1. Validation
2. Verification
3. Data
4. Design

Show Answer

Validation

28. for improving supply chain management to optimize stock management, replenishment, and forecasting

1. Descriptive
2. Diagnostic
3. Predictive
4. Prescriptive

Show Answer

Predictive

29. which among the following is not a Data mining and analytical applications?

1. profile matching
2. social network analysis
3. facial recognition
4. Filtering

Show Answer

Filtering

30. _____ as a result of data accessibility, data latency, data availability, or limits on bandwidth in relation to the size of inputs

Data Analytics Mcq with Answers

1. Computation-restricted throttling
2. Large data volumes
3. Data throttling
4. Data Parallelization

Show Answer

Data throttling

data analytics mcq with answers

31. 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

Show Answer

Increasing revenues

32. Which storage subsystem can support massive data volumes of increasing size.

1. Extensibility
2. Fault tolerance
3. Scalability
4. High-speed I/O capacity

Show Answer

Scalability

33. _____ provides performance through distribution of data and fault tolerance through replication

Data Analytics Mcq with Answers

1. HDFS
2. PIG
3. HIVE
4. HADOOP

Show Answer

HDFS

34. _____ 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

Show Answer

MAP REDUCE

data analytics multiple choice questions

35. _____ 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

Show Answer

REDUCER

36. _____ is a type of local Reducer that groups similar data from the map phase into identifiable sets.

Data Analytics Mcq with Answers

1. MAPPER
2. REDUCER
3. COMBINER
4. PARTITIONER

Show Answer

COMBINER

37. 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

Show Answer

core-site.xml

38. Movie Recommendation systems are an example of

1. **Classification**
 2. **Clustering**
 3. **Reinforcement Learning**
 4. **Regression**
1. 2 only
 2. 1 and 3
 3. 1 and 2
 4. 2 and 3

Show Answer

1 and 3

Data Analytics Mcq with Answers

39. Sentiment Analysis is an example of

1. Regression
 2. Classification
 3. clustering
 4. Reinforcement Learning
-
1. 1, 2 and 4
 2. 1, 2 and 3
 3. 1 and 3
 4. 1 and 2

Show Answer

1, 2 and 4

Unit 4-Classification

1. 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.

a) Decision tree

b) Graphs

c) Trees

d) Neural Networks

2. Decision Tree is a display of an algorithm.

a) True

b) False

3. What is Decision Tree?

a) Flow-Chart

b) Structure in which internal node represents test on an attribute, each branch represents outcome of test and each leaf node represents class label

c) 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

d) None of the mentioned

4. Decision Trees can be used for Classification Tasks.

a) True

b) False

Explanation: None.

5. Choose from the following that are Decision Tree nodes?

a) Decision Nodes

b) End Nodes

c) Chance Nodes

d) All of the mentioned

6. Decision Nodes are represented by _____

a) Disks

b) Squares

c) Circles

d) Triangles

7. Chance Nodes are represented by _____

a) Disks

b) Squares

c) Circles

d) Triangles

8. End Nodes are represented by _____

- a) Disks
- b) Squares
- c) Circles
- d) Triangles**

9. Which of the following are the advantage/s of Decision Trees?

- a) Possible Scenarios can be added
- b) Use a white box model, If given result is provided by a model
- c) Worst, best and expected values can be determined for different scenarios
- d) All of the mentioned**

10. How many terms are required for building a bayes model?

- a) 1
- b) 2
- c) 3**
- d) 4

Explanation: The three required terms are a conditional probability and two unconditional probability.

11. What is needed to make probabilistic systems feasible in the world?

- a) Reliability
- b) Crucial robustness**
- c) Feasibility
- d) None of the mentioned

Explanation: On a model-based knowledge provides the crucial robustness needed to make probabilistic system feasible in the real world.

12. Where does the bayes rule can be used?

- a) Solving queries
- b) Increasing complexity
- c) Decreasing complexity
- d) Answering probabilistic query**

Explanation: Bayes rule can be used to answer the probabilistic queries conditioned on one piece of evidence.

13. What does the bayesian network provides?

- a) Complete description of the domain**
- b) Partial description of the domain
- c) Complete description of the problem
- d) None of the mentioned

Explanation: A Bayesian network provides a complete description of the domain.

14. How the entries in the full joint probability distribution can be calculated?

- a) Using variables

b) Using information

- c) Both Using variables & information
- d) None of the mentioned

Explanation: Every entry in the full joint probability distribution can be calculated from the information in the network.

15. How the bayesian network can be used to answer any query?

- a) Full distribution
- b) Joint distribution**
- c) Partial distribution
- d) All of the mentioned

Explanation: If a bayesian network is a representation of the joint distribution, then it can solve any query, by summing all the relevant joint entries.

16. How the compactness of the bayesian network can be described?

- a) Locally structured**
- b) Fully structured
- c) Partial structure
- d) All of the mentioned

Explanation: The compactness of the bayesian network is an example of a very general property of a locally structured system.

17. To which does the local structure is associated?

- a) Hybrid
- b) Dependant
- c) Linear**
- d) None of the mentioned

Explanation: Local structure is usually associated with linear rather than exponential growth in complexity.

18. Which condition is used to influence a variable directly by all the others?

- a) Partially connected
- b) Fully connected**
- c) Local connected
- d) None of the mentioned

19. What is the consequence between a node and its predecessors while creating bayesian network?

- a) Functionally dependent
- b) Dependant
- c) Conditionally independent**

d) Both Conditionally dependant & Dependant

Explanation: The semantics to derive a method for constructing bayesian networks were led to the consequence that a node can be conditionally independent of its predecessors.

20.How to represent Chance Nodes?

- (A). Disks
- (B). Squares
- (C). Circles**
- (D). Triangles

21.How to represent End Nodes?

- (A). Disks
- (B). Squares
- (C). Circles
- (D). Triangles**

22.Which of the following is a Decision Tree?

- (A). Flow-Chart
- (B). Structure in which internal node represents test on an attribute, every leaf node represents class label and each branch represents outcome of test
- (C). All of these
- (D). None of these**

23.Which of the following are the pros of Decision Trees?

- (A). Possible Scenarios can be added
- (B). Use a white-box model, If a particular result is provided by a model
- (C). best, Worst and expected values can be determined for different scenarios
- (D). All of these**

MCQ Answer: d

24. Which of the following tool is considered as a decision support tool that uses a tree-like graph or model of decisions and their probable results, including utility, chance event outcomes, and resource costs.

(A). Decision tree

(B). Graphs

(C). Trees

(D). Neural Networks

25. Which of the following nodes are Decision Tree nodes?

(A). Decision Nodes

(B). End Nodes

(C). Chance Nodes

(D). All of these

26. Decision Tree is a display of an algorithm.

(A). True

(B). False

(C). Partially true

27. We can use Decision Trees for Classification Tasks.

(A). True

(B). False

(C). Partially true

28. How to represent Decision Nodes?

(A). Disks

(B). Squares

(C). Circles

(D). Triangles

(E). None of these

29. Which of the following is the consequence between a node and its predecessors while creating bayesian network?

(A) Conditionally independent

- (B) Functionally dependent
- (C) Both Conditionally dependant & Dependant
- (D) Dependent

30. Bayes rule can be used for:-

- (A) Solving queries
- (B) Increasing complexity
- (C) Answering probabilistic query**
- (D) Decreasing complexity

31. _____ provides way and means of weighing up the desirability of goals and the likelihood of achieving them.

- (A) Utility theory**
- (B) Decision theory
- (C) Bayesian networks
- (D) Probability theory

32. Which of the following provided by the Bayesian Network?

- (A) Complete description of the problem
- (B) Partial description of the domain
- (C) Complete description of the domain**
- (D) All of the above

33. The entries in the full joint probability distribution can be calculated as

- (A) Using variables
- (B) Both Using variables & information
- (C) Using information**
- (D) All of the above

34. Causal chain (For example, Smoking cause cancer) gives rise to:-

- (A) Conditionally Independence**
- (B) Conditionally Dependence
- (C) Both
- (D) None of the above

35. The bayesian network can be used to answer any query by using:-

- (A) Full distribution
- (B) Joint distribution**
- (C) Partial distribution
- (D) All of the above

UNIT 5-BIG DATA VISUALIZATION

1. 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**

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

2. Data can be visualized using?

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

Explanation: Data visualization is a graphical representation of quantitative information and data by using visual elements like graphs, charts, and maps.

3. 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

Explanation: Data visualization is also an element of the broader data presentation architecture (DPA) discipline, which aims to identify, locate, manipulate, format and deliver data in the most efficient way possible.

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

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

Explanation: Treemaps are best used when multiple categories are present, and the goal is to compare different parts of a whole.

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

- A. fisher.test()
- B. chisq.test()
- C. Lm.test()
- D. prop.test()**

Explanation: prop.test() is used to inference for 1 proportion using normal approx.

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

- A. factor.mosaicplot
- B. factor.xyplot
- C. factor.congruence**

Explanation: factor.congruence is used to find the factor congruence coefficients.

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

- A. qqline()**
- B. qline()
- C. anova()
- D. lm()

Explanation: qqnorm is another tool for checking normality.

8. 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

Explanation: Data visualization decrease the insights andtake solwer decisions is false statement.

9. Common use cases for data visualization include?

- A. Politics
- B. Sales and marketing
- C. Healthcare
- D. All of the above**

Explanation: All option are Common use cases for data visualization.

10. 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

Explanation: If the time series is random, such autocorrelations should be near zero for any and all time-lag separations.

11. Which are pros of data visualization?

- A. It can be accessed quickly by a wider audience.**
- B. It can misrepresent information
- C. It can be distracting
- D. None Of the above

Explanation: Pros of data visualization : it can be accessed quickly by a wider audience.

12. Which are cons of data visualization?

- A. It conveys a lot of information in a small space.
- B. It makes your report more visually appealing.
- C. visual data is distorted or excessively used.**
- D. None Of the above

Explanation: It can be distracting : if the visual data is distorted or excessively used.

13. Which of the intricate techniques is not used for data visualization?

- A. Bullet Graphs
- B. Bubble Clouds
- C. Fever Maps**
- D. Heat Maps

Explanation: Fever Maps is not used for data visualization instead of that Fever charts is used.

14. Which one of the following is most basic and commonly used techniques?

- A. Line charts**
- B. Scatter plots
- C. Population pyramids
- D. Area charts

Explanation: Line charts. This is one of the most basic and common techniques used. Line charts display how variables can change over time.

15. Which is used to query and edit graphical settings?

- A. anova()
- B. par()**
- C. plot()
- D. cum()

Explanation: par() is used to query and edit graphical settings.

16. Which of the following method make vector of repeated values?

- A. rep()
- B. data()**
- C. view()
- D. read()

Explanation: data() load (often into a data.frame) built-in dataset.

17. Who calls the lower level functions lm.fit?

- A. lm()**
- B. col.max
- C. par
- D. histo

Explanation: lm calls the lower level functions lm.fit.

18. Which of the following lists names of variables in a data.frame?

- A. par()
- B. names()
- C. barchart()
- D. quantile()**

Explanation: names function is used to associate name with the value in the vector.

19. Which of the following statement is true?

- A. Scientific visualization, sometimes referred to in shorthand as SciVis
- B. Healthcare professionals frequently use choropleth maps to visualize important health data.
- C. Candlestick charts are used as trading tools and help finance professionals analyze price movements over time
- D. All of the above**

Explanation: All options are correct.

20. Which of the following adds marginal sums to an existing table?

- a) par()
- b) prop.table()**
- c) addmargins()
- d) quantile()

Explanation: prop.table() computes proportions from a contingency table.

21. Which of the following lists names of variables in a data.frame?

- a) quantile()**
- b) names()
- c) barchart()
- d) par()

Explanation: names function is used to associate name with the value in the vector.

22. Which of the following is tool for chi-square distributions?

- a) pchisq()
- b) chisq()
- c) pnorm**
- d) barchart()

Explanation: pnorm() is tool for normal distributions.

23. Which of the following groups values of a variable into larger bins?

- a) cut**
- b) col.max(x)
- c) stem
- d) which.max(x)

Explanation: stem() is used to make a stemplot.

24. Which of the following determine the least-squares regression line?

- a) `histo()`
- b) `lm`**
- c) `barlm()`
- d) `col.max(x)`

Explanation: `lm` calls the lower level functions `lm.fit`.

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

- a) `qqline()`**
- b) `qline()`
- c) `anova()`
- d) `lm()`

Explanation: `qqnorm` is another tool for checking normality.

26. Which of the following is lattice command for producing boxplots?

- a) `plot()`
- b) `bwplot()`**
- c) `xyplot()`
- d) `barlm()`

Explanation: The function `bwplot()` makes box-and-whisker plots for numerical variables.

27. Which of the following compute analysis of variance table for fitted model?

- a) `ecdf()`
- b) `cum()`
- c) `anova()`**
- d) `bwplot()`

Explanation: `ecdf()` builds empirical cumulative distribution function.

28. Which of the following is used to find variance of all values?

- a) `var()`**
- b) `sd()`
- c) `mean()`
- d) `anova()`

Explanation: `sd()` is used to calculate standard deviation.

30. The purpose of `fisher.test()` is _____ test for contingency table.

- a) Chisq
- b) Fisher**
- c) Prop

d) Stem

Explanation: `prop.test()` is used to inference for 1 proportion using normal approx.

UNIT 3

1. A set of activities that ensure that software correctly implements a specific function.

- a) **verification**
- b) testing
- c) implementation
- d) validation

Explanation: Verification ensures that software correctly implements a specific function. It is a static practice of verifying documents.

2. Validation is computer based.

- a) **True**
- b) False

Explanation: The statement is true. Validation is a computer based process. It uses methods like black box testing, gray box testing, etc.

3. _____ is done in the development phase by the debuggers.

- a) Coding
- b) Testing
- c) **Debugging**
- d) Implementation

Explanation: Coding is done by the developers. In debugging, the developer fixes the bug in the development phase. Testing is conducted by the testers.

4. Locating or identifying the bugs is known as _____

- a) Design
- b) **Testing**
- c) Debugging
- d) Coding

Explanation: Testing is conducted by the testers. They locate or identify the bugs. In debugging developer fixes the bug. Coding is done by the developers.

5. Which defines the role of software?

- a) System design
- b) Design
- c) **System engineering**
- d) Implementation

Explanation: The answer is system engineering. System engineering defines the role of software.

6. What do you call testing individual components?

- a) system testing
- b) **unit testing**
- c) validation testing

d) black box testing

Explanation: The testing strategy is called unit testing. It ensures a function properly works as a unit.

7. A testing strategy that tests the application as a whole.

- a) Requirement Gathering
- b) Verification testing
- c) Validation testing
- d) System testing**

Explanation: Validation testing tests the application as a whole against the user requirements. In system testing, it tests the application in the context of an entire system.

8. A testing strategy that tests the application in the context of an entire system.

- a) System**
- b) Validation
- c) Unit
- d) Gray box

Explanation: In system testing, it tests the application in the context of an entire system. The software and other system elements are tested as a whole.

9. A _____ is tested to ensure that information properly flows into and out of the system.

- a) module interface**
- b) local data structure
- c) boundary conditions
- d) paths

Explanation: A module interface is tested to ensure that information properly flows into and out of the system.

10. A testing conducted at the developer's site under validation testing.

- a) alpha**
- b) gamma
- c) lambda
- d) unit

Explanation: Alpha testing is conducted at developer's site. It is conducted by customer in developer's presence before software delivery.

11. In practice, Line of best fit or regression line is found when _____

- a) Sum of residuals ($\sum(Y - h(X))$) is minimum
- b) Sum of the absolute value of residuals ($\sum|Y-h(X)|$) is maximum
- c) Sum of the square of residuals ($\sum(Y-h(X))^2$) is minimum**
- d) Sum of the square of residuals ($\sum(Y-h(X))^2$) is maximum

Explanation: Here we penalize higher error value much more as compared to the smaller one, such that there is a significant difference between making big errors and small errors, which makes it easy to differentiate and select the best fit line.

12. If Linear regression model perfectly fits i.e., train error is zero, then _____

- a) Test error is also always zero

- b) Test error is non zero
- c) Couldn't comment on Test error**
- d) Test error is equal to Train error

Explanation: Test Error depends on the test data. If the Test data is an exact representation of train data then test error is always zero. But this may not be the case.

13. Which of the following metrics can be used for evaluating regression models?

- i) R Squared
 - ii) Adjusted R Squared
 - iii) F Statistics
 - iv) RMSE / MSE / MAE
- a) ii and iv
 - b) i and ii
 - c) ii, iii and iv
 - d) i, ii, iii and iv**

Explanation: These (R Squared, Adjusted R Squared, F Statistics, RMSE / MSE / MAE) are some metrics which you can use to evaluate your regression model.

14. How many coefficients do you need to estimate in a simple linear regression model (One independent variable)?

- a) 1
- b) 2**
- c) 3
- d) 4

Explanation: In simple linear regression, there is one independent variable so 2 coefficients ($Y=a+bx+\text{error}$).

15. In a simple linear regression model (One independent variable), If we change the input variable by 1 unit. How much output variable will change?

- a) by 1
- b) no change
- c) by intercept
- d) by its slope**

Explanation: For linear regression $Y=a+bx+\text{error}$. If neglect error then $Y=a+bx$. If x increases by 1, then $Y = a+b(x+1)$ which implies $Y=a+bx+b$. So Y increases by its slope.

16. Function used for linear regression in R is _____

- a) lm(formula, data)**
- b) lr(formula, data)
- c) lrm(formula, data)

d) `regression.linear(formula, data)`

Explanation: `lm(formula, data)` refers to a linear model in which formula is the object of the class “formula”, representing the relation between variables. Now this formula is applied on the data to create a relationship model.

17. In syntax of linear model `lm(formula,data,...)`, data refers to _____

- a) Matrix
- b) Vector**
- c) Array
- d) List

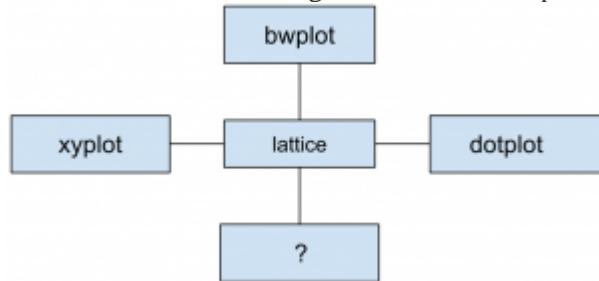
Explanation: Formula is just a symbol to show the relationship and is applied on data which is a vector. In General, `data.frame` are used for data.

18. In the mathematical Equation of Linear Regression $Y = \beta_1 + \beta_2X + \epsilon$, (β_1, β_2) refers to _____

- a) (X-intercept, Slope)
- b) (Slope, X-Intercept)
- c) (Y-Intercept, Slope)**
- d) (slope, Y-Intercept)

Explanation: Y-intercept is β_1 and X-intercept is $-(\beta_1 / \beta_2)$. Intercepts are defined for axis and formed when the coordinates are on the axis.

19. Which of the following function can be replaced with the question mark in the below figure?



- a) boxplot
- b) lplot
- c) levelplot**
- d) all of the mentioned

Explanation: `levelplot` is used plotting “image”.

20. Point out the correct statement.

- a) The mean is a measure of central tendency of the data
- b) Empirical mean is related to “centering” the random variables
- c) The empirical standard deviation is a measure of spread
- d) All of the mentioned**

Explanation: The process of centering and scaling the data is called “normalizing” the data.

21. Which of the following implies no relationship with respect to correlation?

- a) $\text{Cor}(X, Y) = 1$
- b) $\text{Cor}(X, Y) = 0$
- c) $\text{Cor}(X, Y) = 2$
- d) All of the mentioned

Explanation: Correlation is a statistical technique that can show whether and how strongly pairs of variables are related.

22. Normalized data are centered at ___ and have units equal to standard deviations of the original data.

- a) **0**
- b) 5
- c) 1
- d) 10

Explanation: In statistics and applications of statistics, normalization can have a range of meanings.

23. Point out the wrong statement.

- a) Regression through the origin yields an equivalent slope if you center the data first
- b) Normalizing variables results in the slope being the correlation
- c) **Least squares is not an estimation tool**
- d) None of the mentioned

Explanation: Least squares is an estimation tool.

24. Which of the following is correct with respect to residuals?

- a) Positive residuals are above the line, negative residuals are below
- b) **Positive residuals are below the line, negative residuals are above**
- c) Positive residuals and negative residuals are below the line

29. Minimizing the likelihood is the same as maximizing -2 log likelihood.

- a) **True**
- b) False

Explanation: Maximizing the likelihood is the same as minimizing 2 log likelihood.

30. Residuals are useful for investigating best model fit.

- a) True
- b) **False**

Explanation: Residuals are useful for investigating poor model fit.

UNIT 2 BASIC DATA ANALYTIC METHODS

Statistical Methods for Evaluation- Hypothesis testing, difference of means, wilcoxon rank-sum test, type 1 type 2 errors, power and sample size, ANNOVA. Advanced Analytical Theory and Methods: Clustering- Overview, K means- Use cases, Overview of methods, determining number of clusters, diagnostics, reasons to choose and cautions.

1. A statement made about a population for testing purpose is called?

- a)Statistic
- b)Hypothesis**
- c)Level of Significance
- d)Test-Statistic

Explanation: Hypothesis is a statement made about a population in general. It is then tested and correspondingly accepted if True and rejected if False.

2. If the assumed hypothesis is tested for rejection considering it to be true is called?

- a)Null Hypothesis**
- b)Statistical Hypothesis
- c)Simple Hypothesis
- d)Composite Hypothesis

Explanation: If the assumed hypothesis is tested for rejection considering it to be true is called Null Hypothesis. It gives the value of population parameter.

3. A statement whose validity is tested on the basis of a sample is called?

- a)Null Hypothesis
- b)Statistical Hypothesis**
- c)Simple Hypothesis
- d)Composite Hypothesis

Explanation: In testing of Hypothesis a statement whose validity is tested on the basis of a sample is called as Statistical Hypothesis. Its validity is tested with respect to a sample.

4. A hypothesis which defines the population distribution is called?

- a)Null Hypothesis
- b)Statistical Hypothesis
- c)Simple Hypothesis**
- d)Composite Hypothesis

Explanation: A hypothesis which defines the population distribution is called as Simple hypothesis. It specifies all parameter values.

5. If the null hypothesis is false then which of the following is accepted?

- a)Null Hypothesis
- b)Positive Hypothesis
- c)Negative Hypothesis

d) Alternative Hypothesis.

Explanation: If the null hypothesis is false then Alternative Hypothesis is accepted. It is also called as Research Hypothesis.

6. The rejection probability of Null Hypothesis when it is true is called as?

- a) Level of Confidence
- b) Level of Significance**
- c) Level of Margin
- d) Level of Rejection

Answer:b

Explanation: Level of Significance is defined as the probability of rejection of a True Null Hypothesis. Below this probability a Null Hypothesis is rejected.

7. The point where the Null Hypothesis gets rejected is called as?

- a) Significant Value
- b) Rejection Value
- c) Acceptance Value
- d) Critical Value**

Answer:d

Explanation: The point where the Null Hypothesis gets rejected is called as Critical Value. It is also called as dividing point for separation of the regions where hypothesis is accepted and rejected.

8. If the Critical region is evenly distributed then the test is referred as?

- a) Two tailed**
- b) One tailed
- c) Three tailed
- d) Zero tailed

Answer:

a

Explanation: In two tailed test the Critical region is evenly distributed. One region contains the area where Null Hypothesis is accepted and another contains the area where it is rejected.

9. The type of test is defined by which of the following?

- a) Null Hypothesis
- b) Simple Hypothesis
- c) Alternative Hypothesis**
- d) Composite Hypothesis

Answer:c

Explanation: Alternative Hypothesis defines whether the test is one tailed or two tailed. It is also called as Research Hypothesis.

10. Which of the following is defined as the rule or formula to test a Null Hypothesis?

- a) Test statistic**
- b) Population statistic
- c) Variance statistic
- d) Null statistic

Explanation: Test statistic provides a basis for testing a Null Hypothesis. A test statistic is a random variable that is calculated from sample data and used in a hypothesis test.

13. Type I error occurs when?

- a) We reject H_0 if it is True
- b) We reject H_0 if it is False
- c) We accept H_0 if it is True
- d) We accept H_0 if it is False

Explanation: In Testing of Hypothesis Type 1 error occurs when we reject H_0 if it is True. On the contrary a Type 2 error occurs when we accept H_0 if it is False.

14. The probability of Type 1 error is referred as?

- a) $1-\alpha$
- b) β
- c) α
- d) $1-\beta$

Explanation: In Testing of Hypothesis Type 1 error occurs when we reject H_0 if it is True. The probability of H_0 is α then the error probability will be $1-\alpha$.

15. Alternative Hypothesis is also called as?

- a) Composite hypothesis
- b) Research Hypothesis
- c) Simple Hypothesis
- d) Null Hypothesis

Explanation: Alternative Hypothesis is also called as Research Hypothesis. If the Null Hypothesis is false then Alternative Hypothesis is accepted.

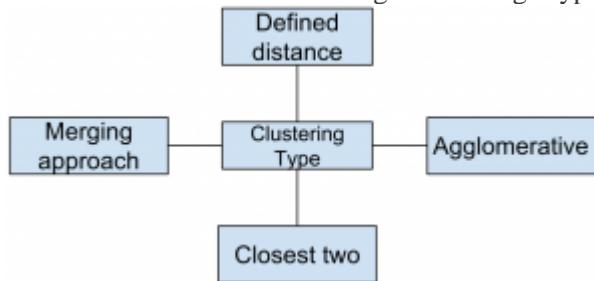
16. In hypothesis testing, a Type 2 error occurs when

- A. The null hypothesis is not rejected when the null hypothesis is true.
- B. The null hypothesis is rejected when the null hypothesis is true.
- C. **The null hypothesis is not rejected when the alternative hypothesis is true.**
- D. The null hypothesis is rejected when the alternative hypothesis is true.

17. Null and alternative hypotheses are statements about:

- A. **population parameters.**
- B. sample parameters.
- C. sample statistics.
- D. it depends - sometimes population parameters and sometimes sample statistics.

18. Which of the following clustering type has characteristic shown in the below figure?



- a) Partitional
- b) Hierarchical**
- c) Naive bayes
- d) None of the mentioned

Explanation: Hierarchical clustering groups data over a variety of scales by creating a cluster tree or dendrogram.

19. Point out the correct statement.

- a) The choice of an appropriate metric will influence the shape of the clusters
- b) Hierarchical clustering is also called HCA
- c) In general, the merges and splits are determined in a greedy manner
- d) All of the mentioned**

Explanation: Some elements may be close to one another according to one distance and farther away according to another.

20. Which of the following is finally produced by Hierarchical Clustering?

- a) final estimate of cluster centroids
- b) tree showing how close things are to each other**
- c) assignment of each point to clusters
- d) all of the mentioned

Explanation: Hierarchical clustering is an agglomerative approach.

21. Which of the following is required by K-means clustering?

- a) defined distance metric
- b) number of clusters
- c) initial guess as to cluster centroids
- d) all of the mentioned**

Explanation: K-means clustering follows partitioning approach.

22. Point out the wrong statement.

- a) k-means clustering is a method of vector quantization
- b) k-means clustering aims to partition n observations into k clusters
- c) k-nearest neighbor is same as k-means**
- d) none of the mentioned

Explanation: k-nearest neighbor has nothing to do with k-means.

23. Which of the following combination is incorrect?

- a) Continuous – euclidean distance
- b) Continuous – correlation similarity

c) Binary – manhattan distance

d) None of the mentioned

Explanation: You should choose a distance/similarity that makes sense for your problem.

24. Hierarchical clustering should be primarily used for exploration.

a) True

b) False

Explanation: Hierarchical clustering is deterministic.

25. Which of the following function is used for k-means clustering?

a) k-means

b) k-mean

c) heat map

d) none of the mentioned

Explanation: K-means requires a number of clusters.

26. Which of the following clustering requires merging approach?

a) Partitional

b) Hierarchical

c) Naive Bayes

d) None of the mentioned

Explanation: Hierarchical clustering requires a defined distance as well.

27. K-means is not deterministic and it also consists of number of iterations.

a) True

b) False

Explanation: K-means clustering produces the final estimate of cluster centroids.

28. For two runs of K-Mean clustering is it expected to get same clustering results?

A. Yes

B. No

29. Is it possible that Assignment of observations to clusters does not change between successive iterations in K-Means

A. Yes

B. No

C. Can't say

D. None of these

When the K-Means algorithm has reached the local or global minima, it will not alter the assignment of data points to clusters for two successive iterations.

30. 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.

Options:

A. 1, 3 and 4

B. 1, 2 and 3

C. 1, 2 and 4

D. All of the above

All four conditions can be used as possible termination condition in K-Means clustering:

1. This condition limits the runtime of the clustering algorithm, but in some cases the quality of the clustering will be poor because of an insufficient number of iterations.
2. Except for cases with a bad local minimum, this produces a good clustering, but runtimes may be unacceptably long.
3. This also ensures that the algorithm has converged at the minima.
4. Terminate when RSS falls below a threshold. This criterion ensures that the clustering is of a desired quality after termination. Practically, it's a good practice to combine it with a bound on the number of iterations to guarantee termination.

31. Which of the following clustering algorithms suffers from the problem of convergence at local optima?

1. K- Means clustering algorithm
2. Agglomerative clustering algorithm
3. Expectation-Maximization clustering algorithm
4. Diverse clustering algorithm

Options:

A. 1 only

B. 2 and 3

C. 2 and 4

D. 1 and 3

E. 1,2 and 4

F. All of the above

Out of the options given, only K-Means clustering algorithm and EM clustering algorithm has the drawback of converging at local minima.

32. Which of the following algorithm is most sensitive to outliers?

A. K-means clustering algorithm

B. K-medians clustering algorithm

C. K-modes clustering algorithm

D. K-medoids clustering algorithm

Out of all the options, K-Means clustering algorithm is most sensitive to outliers as it uses the mean of cluster data points to find the cluster center.

UNIT ONE	SUB : 410243 DA					
Sr. No.	Questions	a	b	c	d	Ans
1	Business intelligence (BI) is a broad category of application programs which includes _____	a) Decision support	b) Data mining	c) OLAP	d) All of the mentioned	d
2	BI can catalyze a business's success in terms of _____	a) Distinguish the products and services that drive revenues	b) Rank customers and locations based on profitability	c) Ranks customers and locations based on probability	d) All of the mentioned	d
3	Which of the following areas are affected by BI?	a) Revenue	b) CRM	c) Sales	d) All of the mentioned	b
4	_____ is a performance management tool that recapitulates an organization's performance from several standpoints on a single page.	a) Balanced Scorecard	b) Data Cube	c) Dashboard	d) All of the mentioned	a
5	_____ is a system where operations like data extraction, transformation and loading operations are executed.	a) Data staging	b) Data integration	c) ETL	d) None of the mentioned	a
6	_____ is a category of applications and technologies for presenting and analyzing corporate and external data.	a) Data warehouse	b) MIS	c) EIS	d) All of the mentioned	c
7	Which of the following is the process of basing an organization's actions and decisions on actual measured results of performance?	a) Institutional performance management	b) Gap analysis	c) Slice and Dice	d) None of the mentioned	a

8	Which of the following does not form part of BI Stack in SQL Server?	a) SSRS	b) SSIS	c) SSAS	d) OBIEE	d
9	BI can catalyze a business's success in terms of _____	a) Distinguish the products and services that drive revenues	b) Rank customers and locations based on profitability	c) Ranks customers and locations based on probability	d) All of the mentioned	d
10	This is an approach to selling goods and services in which a prospect explicitly agrees in advance to receive marketing information.	A. customer managed relationship	B. data mining	C. permission marketing	D. one-to-one marketing	c
11	In an Internet context, this is the practice of tailoring Web pages to individual users' characteristics or preferences.	a. Web services	b. customer-facing	c. client/server	d. personalization	d
12	This is the processing of data about customers and their relationship with the enterprise in order to improve the enterprise's future sales and service and lower cost.	a. clickstream analysis	b. database marketing	c. customer relationship management	d. CRM analytics	d
13	This is a broad category of applications and technologies for gathering, storing, analyzing, and providing access to data to help enterprise users make better business decisions.	a. best practice	b. data mart	c. business information warehouse	d. business intelligence	d

14	This is a systematic approach to the gathering, consolidation, and processing of consumer data (both for customers and potential customers) that is maintained in a company's databases.	a. database marketing	b. marketing encyclopedia	c. application integration	d. service oriented integration	a
15	This is an arrangement in which a company outsources some or all of its customer relationship management functions to an application service provider (ASP).	a. spend management	b. supplier relationship management	c. hosted CRM	d. Customer Information Control System	c
16	This is an XML-based metalanguage developed by the Business Process Management Initiative (BPMI) as a means of modeling business processes, much as XML is, itself, a metalanguage with the ability to model enterprise data.	a. BizTalk	b. BPML	c. e-biz	d. ebXML	b
17	This is a central point in an enterprise from which all customer contacts are managed.	a. contact center	b. help system	c. multichannel marketing	d. call center	a
18	This is the practice of dividing a customer base into groups of individuals that are similar in specific ways relevant to marketing, such as age, gender, interests, spending habits, and so on.	a. customer service chat	b. customer managed relationship	c. customer life cycle	d. customer segmentation	d

19	In data mining, this is a technique used to predict future behavior and anticipate the consequences of change.	a. predictive technology	b. disaster recovery	c. phase change	d. predictive modeling	d
20	1. According to analysts, for what can traditional IT systems provide a foundation when they're integrated with big data technologies like Hadoop?	Big data management and data mining	Data warehousing and business intelligence	Management of Hadoop clusters	Collecting and storing unstructured data	a
21	All of the following accurately describe Hadoop, EXCEPT:	Open source	Real-time	Java-based	Distributed computing approach	b
22	_____ has the world's largest Hadoop cluster.	Apple	Datamatics	Facebook	None of the mentioned	c
23	What are the five V's of Big Data?	Volume	velocity	Variety	All of the above	d
24	_____ hides the limitations of Java behind a powerful and concise Clojure API for Cascading.	Scalding	Cascalog	Hcatalog	Hcalding	b
25	What are the main components of Big Data?	MapReduce	HDFS	YARN	All of these	d
26	What are the different features of Big Data Analytics?	Open-Source	Scalability	Data Recovery	All the above	d
27	Define the Port Numbers for NameNode, Task Tracker and Job Tracker.	NameNode	Task Tracker	Job Tracker	All of the above	d

28	Facebook Tackles Big Data With _____ based on Hadoop	Project Prism	Prism	ProjectData	ProjectBid	a
29	What is a unit of data that flows through a Flume agent?	Record	Event	Row	Log	b
30	A feature F1 can take certain value: A, B, C, D, E, & F and represents grade of students from a college. Which of the following statement is true in the following case	Feature F1 is an example of nominal variable.	Feature F1 is an example of ordinal variable.	It doesn't belong to any of the above category.	Both of these	b
31	Which of the following is an example of a deterministic algorithm?	PCA	K-Means	None of the above	all of the above	a
32	What is the entropy of the target variable?	$-(5/8 \log(5/8) + 3/8 \log(3/8))$	$5/8 \log(5/8) + 3/8 \log(3/8)$	$5/8 \log(5/8) + 3/8 \log(3/8)$	$5/8 \log(3/8) - 3/8 \log(5/8)$	a
33	Point out the correct statement.	a) OLAP is an umbrella term that refers to an assortment of software applications for analyzing an organization's raw data for intelligent decision making	b) Business intelligence equips enterprises to gain business advantage from data	c) BI makes an organization agile thereby giving it a lower edge in today's evolving market condition	None of the mentioned	b

34	BI can catalyze a business's success in terms of _____	a) Distinguish the products and services that drive revenues	b) Rank customers and locations based on profitability	c) Ranks customers and locations based on probability	d) All of the mentioned	d
35	Which of the following areas are affected by BI?	a) Revenue	b) CRM	c) Sales	d) All of the mentioned	b
36	Which of the following does not form part of BI Stack in SQL Server?	a) SSRS	b) SSIS	c) SSAS	d) OBIEE	d
37	BI can catalyze a business's success in terms of _____	a) Distinguish the products and services that drive revenues	b) Rank customers and locations based on profitability	c) Ranks customers and locations based on probability	d) All of the mentioned	d
38	Heuristic is	A set of databases from different vendors, possibly using different database paradigms	An approach to a problem that is not guaranteed to work but performs well in most cases	Information that is hidden in a database and that cannot be recovered by a simple SQL query.	None of these	b
39	In an Internet context, this is the practice of tailoring Web pages to individual users' characteristics or preferences.	a. Web services	b. customer-facing	c. client/server	d. personalization	d

40	Heterogeneous databases referred to	A set of databases from different vendors, possibly using different database paradigms	An approach to a problem that is not guaranteed to work but performs well in most cases.	Information that is hidden in a database and that cannot be recovered by a simple SQL query.	None of these	a
----	-------------------------------------	--	--	--	---------------	----------

UNIT TWO	SUB : 410243 DA					
Sr. No.	Questions	a	b	c	d	Ans
1	Movie Recommendation systems are an example of:	Classification	Clustering	Reinforcement Learning	Regression	b,c
2	Sentiment Analysis is an example of:	Regression	Classification	Clustering	Reinforcement Learning	a,b,d
3	What is the minimum no. of variables/ features required to perform clustering?	0	1	2	3	b
4	Is it possible that Assignment of observations to clusters does not change between successive iterations in K-Means	Yes	No	Can't say	None of these	a
5	Which of the following can act as possible termination conditions in K-Means?	For a fixed number of iterations.	Assignment of observations to clusters does not change between iterations. Except for cases with a bad local minimum.	Centroids do not change between successive iterations.	Terminate when RSS falls below a threshold.	a,b,c,d

6	Which of the following clustering algorithms suffers from the problem of convergence at local optima?	K- Means clustering algorithm	Agglomerative clustering algorithm	Expectation-Maximization clustering algorithm	Diverse clustering algorithm	a,c
7	Which of the following algorithm is most sensitive to outliers?	K-means clustering algorithm	K-medians clustering algorithm	K-modes clustering algorithm	K-medoids clustering algorithm	a
8	How can Clustering (Unsupervised Learning) be used to improve the accuracy of Linear Regression model (Supervised Learning):	Creating different models for different cluster groups.	Creating an input feature for cluster ids as an ordinal variable.	Creating an input feature for cluster centroids as a continuous variable.	Creating an input feature for cluster size as a continuous variable.	a,b,c,d
9	What could be the possible reason(s) for producing two different dendograms using agglomerative clustering algorithm for the same dataset?	Proximity function used	of data points used	of variables used	All of the above	d
10	In which of the following cases will K-Means clustering fail to give good results?	Data points with outliers	Data points with different densities	Data points with round shapes	Data points with non-convex shapes	a,b,d
11	Which of the following is/are valid iterative strategy for treating missing values before clustering analysis?	Imputation with mean	Nearest Neighbor assignment	Imputation with Expectation Maximization algorithm	All of the above	c

12	Feature scaling is an important step before applying K-Mean algorithm. What is reason behind this?	In distance calculation it will give the same weights for all features	You always get the same clusters. If you use or don't use feature scaling	In Manhattan distance it is an important step but in Euclidian it is not	None of these	a
13	Which of the following method is used for finding optimal of cluster in K-Mean algorithm?	Elbow method	Manhattan method	Ecludian mehthod	All of the above	a
14	What is true about K-Mean Clustering?	K-means is extremely sensitive to cluster center initializations	Bad initialization can lead to Poor convergence speed	Bad initialization can lead to bad overall clustering	None of these	d
15	Which of the following can be applied to get good results for K-means algorithm corresponding to global minima?	Try to run algorithm for different centroid initialization	Adjust number of iterations	Find out the optimal number of clusters	None of these	a,b,c
16	If you are using Multinomial mixture models with the expectation-maximization algorithm for clustering a set of data points into two clusters, which of the assumptions are important:	All the data points follow two Gaussian distribution	All the data points follow n Gaussian distribution ($n > 2$)	All the data points follow two multinomial distribution	All the data points follow n multinomial distribution ($n > 2$)	c

17	Which of the following is/are not true about Centroid based K-Means clustering algorithm and Distribution based expectation-maximization clustering algorithm:	Both starts with random initializations	Both are iterative algorithms	Both have strong assumptions that the data points must fulfill	Expectation maximization algorithm is a special case of K-Means	d
18	Which of the following is/are not true about DBSCAN clustering algorithm:	For data points to be in a cluster, they must be in a distance threshold to a core point	It has strong assumptions for the distribution of data points in dataspace	It has substantially high time complexity of order O(n ³)	It does not require prior knowledge of the no. of desired clusters	b,c
19	Which of the following are the high and low bounds for the existence of F-Score?	[0,1]	(0,1)	[-1,1]	None of the above	a
20	1. All of the following increase the width of a confidence interval except:	a. Increased confidence level	b. Increased variability	c. Increased sample size	d. Decreased sample size	c
21	3The p-value in hypothesis testing represents which of the following: Please select the best answer of those provided below.	a. The probability of failing to reject the null hypothesis, given the observed results	b. The probability that the null hypothesis is true, given the observed results	c. The probability that the observed results are statistically significant, given that the null hypothesis is true	d. The probability of observing results as extreme or more extreme than currently observed, given that the null hypothesis is true	d

22	4. Assume that the difference between the observed, paired sample values is defined in the same manner and that the specified significance level is the same for both hypothesis tests. Using the same data, the statement that “a paired/dependent two sample t-test is equivalent to a one sample t-test on the paired differences, resulting in the same test statistic, same p-value, and same conclusion” is: Please select the best answer of those provided below.	a. Always True 	b. Never True 	c. Sometimes True 	d. Not Enough Information 	a
23	19. Green sea turtles have normally distributed weights, measured in kilograms, with a mean of 134.5 and a variance of 49.0. A particular green sea turtle’s weight has a z-score of -2.4. What is the weight of this green sea turtle? Round to the nearest whole number.	a. 17 kg 	b. 151 kg 	c. 118 kg 	d. 252 kg 	c
24	What percentage of measurements in a dataset fall above the median?	a. 49% 	b. 50% 	c. 51% 	d. Cannot Be Determined 	d
25	24. The proportion of variation in 5k race times that can be explained by the variation in the age of competitive male runners was approximately 0.663. What is the value of the sample linear correlation coefficient? Round to 3 decimal places.	a. 0.663 	b. 0.814 	c. -0.814 	d. 0.440 	c

26	25. Using all of the results provided, is it reasonable to predict the 5k race time (minutes) of a competitive male runner 73 years of age?	a. Yes; linear correlation between age and 5k race times is statistically significant	b. Yes; both the sample linear regression equation and an age in years is provided	c. No; linear correlation between age and 5k race times is not statistically significant	d. No; the age provided is beyond the scope of our available sample data	d
27	Algorithm is	It uses machine-learning techniques. Here program can learn from past experience and adapt themselves to new situations	Computational procedure that takes some value as input and produces some value as output	Science of making machines performs tasks that would require intelligence when performed by humans	None of these	b

28	Bias is	A class of learning algorithm that tries to find an optimum classification of a set of examples using the probabilistic theory	Any mechanism employed by a learning system to constrain the search space of a hypothesis	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.	None of these	b
29	Classification is	A subdivision of a set of examples into a number of classes	A measure of the accuracy, of the classification of a concept that is given by a certain theory	The task of assigning a classification to a set of examples	None of these	a

30	Binary attribute are	This takes only two values. In general, these values will be 0 and 1 and .they can be coded as one bit	The natural environment of a certain species	Systems that can be used without knowledge of internal operations	None of these	a
31	Classification accuracy is	A subdivision of a set of examples into a number of classes	Measure of the accuracy, of the classification of a concept that is given by a certain theory	The task of assigning a classification to a set of examples	None of these	b
32	Cluster is	Group of similar objects that differ significantly from other objects	Operations on a database to transform or simplify data in order to prepare it for a machine-learning algorithm	Symbolic representation of facts or ideas from which information can potentially be extracted	None of these	a

33	A definition of a concept is----if it recognizes all the instances of that concept	Complete	Consistent	Constant	None of these	a
34	A definition or a concept is----- if it classifies any examples as coming within the concept	Complete	Consistent	Constant	None of these	b
35	Data selection is	The actual discovery phase of a knowledge discovery process	The stage of selecting the right data for a KDD process	A subject-oriented integrated time variant non-volatile collection of data in support of management	None of these	b
36	Classification task referred to	A subdivision of a set of examples into a number of classes	A measure of the accuracy, of the classification of a concept that is given by a certain theory	The task of assigning a classification to a set of examples	None of these	c

37	Hybrid is	Combining different types of method or information	Approach to the design of learning algorithms that is structured along the lines of the theory of evolution.	Decision support systems that contain an information base filled with the knowledge of an expert formulated in terms of if-then rules.	None of these	a
38	Discovery is	It is hidden within a database and can only be recovered if one is given certain clues (an example IS encrypted information).	The process of executing implicit previously unknown and potentially useful information from data	An extremely complex molecule that occurs in human chromosomes and that carries genetic information in the form of genes.	None of these	b
39	What could be the possible reason(s) for producing two different dendograms using agglomerative clustering algorithm for the same dataset?	Proximity function used	of data points used	of variables used	All of the above	d

40	Is it possible that Assignment of observations to clusters does not change between successive iterations in K-Means	Yes	No	Can't say	None of these	a

UNIT THREE	SUB : 410243 DA					
Sr. No.	Questions	a	b	c	d	Ans
1	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	a
2	The correlation coefficient for two real-valued attributes is -0.85. What does this value tell you?	The attributes are not linearly related.	As the value of one attribute decreases the value of the second attribute increases.	As the value of one attribute increases the value of the second attribute also increases.	The attributes show a linear relationship	b
3	Given a rule of the form IF X THEN Y, rule confidence is defined as the conditional probability that	Y is false when X is known to be false.	Y is true when X is known to be true.	X is true when Y is known to be true	X is false when Y is known to be false.	b
4	Chameleon is	Density based clustering algorithm	Partitioning based algorithm	Model based algorithm	Hierarchical clustering algorithm	d
5	Find odd man out	DBSCAN	K-Mean	PAM	None of above	a
6	The number of iterations in apriori _____	increases with the size of the data	decreases with the increase in size of the data	increases with the size of the maximum frequent set	decreases with increase in size of the maximum frequent set	c

7	Which of the following are interestingness measures for association rules?	Recall	Lift	Accuracy	All of Above	b
8	Given a frequent itemset L, If $ L = k$, then there are	$2^k - 1$ candidate association rules	2^k candidate association rules	2^{k-2} candidate association rules	2^{k-2} candidate association rules	c
9	_____ is an example for case based-learning	Decision trees	Neural networks	Genetic algorithm	K-nearest neighbor	d
10	The average positive difference between computed and desired outcome values.	mean positive error	mean squared error	mean absolute error	root mean squared error	c
11	Frequent item sets is	Superset of only closed frequent item sets	Superset of only maximal frequent item sets	Subset of maximal frequent item sets	Superset of both closed frequent item sets and maximal frequent item sets	d
12	Assume that we have a dataset containing information about 200 individuals. A supervised data mining session has discovered the following rule: IF age < 30 & credit card insurance = yes THEN life insurance = yes Rule Accuracy: 70% and Rule Coverage: 63% How many individuals in the class life insurance= no have credit card insurance and are less than 30 years old?	63	38	40	89	b
13	Which of the following is cluster analysis?	Simple segmentation	Grouping similar objects	Labeled classification	Query results grouping	b
14	A good clustering method will produce high quality clusters with	high inter class similarity	high intra class similarity	low intra class similarity	None of above	c

15	Which two parameters are needed for DBSCAN	Min threshold	Min points and eps	Min sup and min confidence	Number of centroids	b
16	Which statement is true about neural network and linear regression models?	Both techniques build models whose output is determined by a linear sum of weighted input attribute values.	The output of both models is a categorical attribute value.	Both models require numeric attributes to range between 0 and 1.	Both models require input attributes to be numeric.	d
17	In Apriori algorithm, if 1 item-sets are 100, then the number of candidate 2 item-sets are	100	200	4950	5000	c
18	Significant Bottleneck in the Apriori algorithm is	Finding frequent itemsets	Pruning	Candidate generation	Number of iterations	c
19	Machine learning techniques differ from statistical techniques in that machine learning methods	are better able to deal with missing and noisy data	typically assume an underlying distribution for the data	have trouble with large-sized datasets	are not able to explain their behavior.	a
20	The probability of a hypothesis before the presentation of evidence.	a priori	posterior	conditional	subjective	a
21	KDD represents extraction of	data	knowledge	rules	model	b

22	Which statement about outliers is true?	Outliers should be part of the training dataset but should not be present in the test data.	Outliers should be identified and removed from a dataset.	The nature of the problem determines how outliers are used	Outliers should be part of the test dataset but should not be present in the training data.	c
23	The most general form of distance is	Manhattan	Eucledian	Mean	Minkowski	d
24	Which Association Rule would you prefer	High support and medium confidence	High support and low confidence	Low support and high confidence	Low support and low confidence	c
25	In a Rule based classifier, If there is a rule for each combination of attribute values, what do you called that rule set R	Exhaustive	Inclusive	Comprehensive	Mutually exclusive	a
26	The apriori property means	If a set cannot pass a test, its supersets will also fail the same test	To decrease the efficiency, do level-wise generation of frequent item sets	To improve the efficiency, do level-wise generation of frequent item sets	If a set can pass a test, its supersets will fail the same test	a
27	If an item set 'XYZ' is a frequent item set, then all subsets of that frequent item set are	Undefined	Not frequent	Frequent	Can not say	c

28	The probability that a person owns a sports car given that they subscribe to automotive magazine is 40%. We also know that 3% of the adult population subscribes to automotive magazine. The probability of a person owning a sports car given that they don't subscribe to automotive magazine is 30%. Use this information to compute the probability that a person subscribes to automotive magazine given that they own a sports car	0.0368	0.0396	0.0389	0.0398	b
29	Simple regression assumes a _____ relationship between the input attribute and output attribute.	quadratic	inverse	linear	reciprocal	c
30	To determine association rules from frequent item sets	Only minimum confidence needed	Neither support nor confidence needed	Both minimum support and confidence are needed	Minimum support is needed	c
31	If {A,B,C,D} is a frequent itemset, candidate rules which is not possible is	C → A	D → ABCD	A → BC	B → ADC	b
32	Which Association Rule would you prefer	High support and low confidence	Low support and high confidence	Low support and low confidence	High support and medium confidence	b
33	Classification rules are extracted from _____	decision tree	root node	branches	siblings	a
34	What does K refers in the K-Means algorithm which is a non-hierarchical clustering approach?	Complexity	Fixed value	No of iterations	. number of clusters	d

35	If Linear regression model perfectly fit i.e., train error is zero, then _____	Test error is also always zero	Test error is non zero	Couldn't comment on Test error	Test error is equal to Train error	c
36	Which of the following metrics can be used for evaluating regression models? i) R Squared ii) Adjusted R Squared iii) F Statistics iv) RMSE/MSE/MAE	ii and iv	i and ii	ii, iii and iv	i, ii, iii and iv	d
37	How many coefficients do you need to estimate in a simple linear regression model (One independent variable)?	1	2	3	4	b
38	In a simple linear regression model (One independent variable), If we change the input variable by 1 unit. How much output variable will change?	by 1	no change	by intercept	by its slope	d
39	In syntax of linear model lm(formula,data,,), data refers to _____	Matrix	array	vector	list	c
40	In the mathematical Equation of Linear Regression $Y = \beta_1 + \beta_2 X + \epsilon$, (β_1, β_2) refers to _____	(X-intercept, Slope)	(Slope, X-Intercept)	(Y-Intercept, Slope)	(slope, Y-Intercept)	c

UNIT FOUR	SUB : 410243 DA					
Sr. No.	Questions	a	b	c	d	Ans
1	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	a
3	What is Decision Tree?	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 & Structure in which internal node represents test on an attribute, each branch represents outcome of test and each leaf node represents class label	None of Above		c
4	Decision Trees can be used for Classification Tasks.	TRUE	FALSE			a
5	Choose from the following that are Decision Tree nodes?	Decision Nodes	End Nodes	Chance Nodes	All of Above	d
6	Decision Nodes are represented by _____	Disks	Squares	Circles	Triangles	b
7	Chance Nodes are represented by _____	Disks	Squares	Circles	Triangles	c

8	End Nodes are represented by _____	Disks	Squares	Circles	Triangles	d
9	Which of the following are the advantage/s of Decision Trees?	Possible Scenarios can be added	Use a white box model, If given result is provided by a model	Worst, best and expected values can be determined for different scenarios	All of Above	d
10	Which of the following statements about Naive Bayes is incorrect?	Attributes are equally important.	Attributes are statistically dependent of one another given the class value.	Attributes are statistically independent of one another given the class value.	Attributes can be nominal or numeric	b
11	Which of the following is not supervised learning?	Clustering	Decision Tree	Linear Regression	Naive Bayesian	a
12	How many terms are required for building a bayes model?	1	2	3	4	c
13	Where does the bayes rule can be used?	Solving queries	Increasing complexity	Decreasing complexity	Answering probabilistic query	d
14	How the bayesian network can be used to answer any query?	Full distribution	Joint distribution	Partial distribution	All of Above	b

			Functionally dependent	Dependant	Conditionally independent	Both Conditionally dependant & Dependant	c
15	What is the consequence between a node and its predecessors while creating bayesian network?		Functionally dependent	Dependant	Conditionally independent	Both Conditionally dependant & Dependant	c
16	Bayesian classifiers is	A class of learning algorithm that tries to find an optimum classification of a set of examples using the probabilistic theory.	Any mechanism employed by a learning system to constrain the search space of a hypothesis	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.	None of these		a

17	Bias is	A class of learning algorithm that tries to find an optimum classification of a set of examples using the probabilistic theory	Any mechanism employed by a learning system to constrain the search space of a hypothesis	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.	None of these	b
18	Background knowledge referred to	Additional acquaintance used by a learning algorithm to facilitate the learning process	A neural network that makes use of a hidden layer	It is a form of automatic learning.	None of these	a

19	Classification accuracy is	A subdivision of a set of examples into a number of classes	A measure of the accuracy, of the classification of a concept that is given by a certain theory	The task of assigning a classification to a set of examples	None of these	b
20	Classification is	A subdivision of a set of examples into a number of classes	A measure of the accuracy, of the classification of a concept that is given by a certain theory	The task of assigning a classification to a set of examples	None of these	a
21	Discovery is	It is hidden within a database and can only be recovered if one is given certain clues (an example IS encrypted information).	The process of executing implicit previously unknown and potentially useful information from data	An extremely complex molecule that occurs in human chromosomes and that carries genetic information in the form of genes.	None of these	b

22	Classification task referred to	A subdivision of a set of examples into a number of classes	A measure of the accuracy, of the classification of a concept that is given by a certain theory	The task of assigning a classification to a set of examples	None of these	c
23	Euclidean distance measure is	A stage of the KDD process in which new data is added to the existing selection.	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	The distance between two points as calculated using the Pythagoras theorem	None of these	c
24	The problem of finding hidden structure in unlabeled data is called	Supervised learning	Unsupervised learning	Reinforcement learning	None of these	b

25	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	Classification	Regression	Clustering	Structural equation modeling	b
26	Discriminating between spam and ham e-mails is a classification task, true or false?	TRUE	FALSE			a
27	which of the following is not involve in data mining?	Knowledge extraction	Data archaeology	Data exploration	Data transformation	d
28	Naive prediction is	A class of learning algorithms that try to derive a Prolog program from examples	A table with n independent attributes can be seen as an n-dimensional space.	A prediction made using an extremely simple method, such as always predicting the same output.	None of these	c
29	Node is	A component of a network	In the context of KDD and data mining, this refers to random errors in a database table.	One of the defining aspects of a data warehouse	None of these	a

30	Prediction is	The result of the application of a theory or a rule in a specific case	One of several possible entries within a database table that is chosen by the designer as the primary means of accessing the data in the table.	Discipline in statistics that studies ways to find the most interesting projections of multi-dimensional spaces.	None of these	a
31	What is the relation between the distance between clusters and the corresponding class discriminability?	proportional	inversely-proportional	no-relation	None of these	a
32	the classification method in which the upper limit of interval is same as of lower class interval is called....	exclusive method	inclusive method	mid point method	None of these	a
33	larger value is 60 and the smallest value is 40 and the number of classes is 5 then the class interval is	20	25	4	15	c
34	summary and presentation of data in tabular form with several non overlapping classes is referred as	nominal distribution	frequency distribution	ordinal distribution	None of these	b
35	the classification method in which the upper and lower limit of interval is also in class interval itself is called....	exclusive method	inclusive method	mid point method	None of these	b

36	Suppose there are 25 base classifiers. Each classifier has error rates of $e = 0.35$. Suppose you are using averaging as ensemble of above 25 classifiers will make a wrong prediction? Note: all classifiers are independent of each other	0.05	0.06	0.07	0.08	b
37	The most widely used metrics and tools to assess a classification model are:	Confusion matrix	Cost-sensitive accuracy	Area under the ROC curve	All of Above	d
38	When performing regression or classification, which of the following is the correct way to preprocess the data?	Normalize the data → PCA → training	PCA → normalize PCA output → training	Normalize the data → PCA → normalize PCA output → training	None of these	a
39	Which of the following is true about Naive Bayes ?	Assumes that all the features in a dataset are equally important	Assumes that all the features in a dataset are independent	both a and b	None of these	c
40	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	1 and 2	2 and 3	1, 2, and 3	1 and 3	c

UNIT FIVE	SUB : 410243 DA					
Sr. No.	Questions	a	b	c	d	Ans
1	Data visualization is realted with...	Pictorial representaions	numerical representation	numerical calculations	None of these	a
2	Which of the following are Use of data visualtization	See context of data	Clear data understanding	finding pattern in data	all of above	d
3	Which of the following statements are true about using visualizations to display a dataset? I. Visualizations are visually appealing, but don't help the viewer understand relationships that exist in the data II. Visualizations like graphs, charts, or visualizations with pictures are useful for conveying information, while tables just filled with text are not useful. III. Patterns that exist in the data can be found more easily by using a visualization	I AND II	II AND III	I AND III	ONLY III	d
4	The plot method on Series and DataFrame is just a simple wrapper around _____	gplt.plot()	plt.plot()	plt.plotgraph()	none of the mentioned	b
5	Point out the correct combination with regards to kind keyword for graph plotting.	'hist' for histogram	'box' for boxplot	'area' for area plots	all of the mentioned	d

6	Which of the following value is provided by kind keyword for barplot?	bar	bar	bar	none of the mentioned	a
7	You can create a scatter plot matrix using the _____ method in pandas.tools.plotting.	sca_matrix	scatter_matrix	DataFrame.plot	all of the mentioned	b
8	Plots may also be adorned with error bars or tables.	True	FALSE	Cannot Tell	All Above	a
9	Which of the following plots are often used for checking randomness in time series?	Autocausation	Autorank	Autocorrelation	none of the mentioned	c
10	_____ plots are used to visually assess the uncertainty of a statistic	Lag	RadViz	Bootstrap	All Above	c
11	Which of the following is not a challenge in Big Data Visualization>?	Velocity	Volume	Version	Variety	c
12	Which of the following is not a problem in Big Data Visualization>?	Visual Noise	Scaled Data	Large image perception	Information Loss	b
13	Which of the following is a problem in Big Data Visualization>?	Structured Data	Scaled Data	Visual Noise	Multiple valued Data	c
14	Which of the candidate is suitable for interactive visualization?	Type of Visual	Cardinality	Size of data	all of above	d
15	Which of the following follows interactive visualization approach?	Zoom+Pan	Focus+Context	Overview+Details	all of above	d
16	Visual Mapping is important for _____	Remapping	Overview+Details	Focus	Context	a
17	Data visualization techniques are:	Scatter Plot	Line Chart	Pie Chart	all of above	d
18	Information Visualization techniques are	Flow Chart	Time Line	DFD	All of above	d

19	Data visualization techniques are:	Flow Chart	Time Line	Pie Chart	None of these	c
20	Information Visualization techniques are	Flow Chart	Line Chart	Pie Chart	None of these	a
21	Data visualization techniques are:	Scatter Plot	Time Line	DFD	None of these	a
22	Information Visualization techniques are	Scatter Plot	Time Line	Bubble Chart	None of these	b
23	Data visualization techniques are:	Histogram	Parallel Coordinates	Time Line	None of these	a
24	Information Visualization techniques are	Semantic Network	Histogram	Area Chart	None of these	a
25	Which of the following is realted term with correlation?	Exponential	U-Shape	Null	All of above	d
26	Data visualization techniques are:	Scatter Plot	Time Line	DFD	None of these	a
27	Coulmn graph is another name for _____	Bar Chart	Scatterplot	Histogram	Area Chart	a
28	Which of the following follows interactive visualization approach?	Zoom+Pan	Focus+Context	Overview+Details	all of above	d
29	information Visualization techniques are	Pie Chart	Scatterplot	Histogram	Area Chart	a
30	Which of the following is category of timeline?	Linear Timeline	Modular Timeline	Variant Timeline	ER Timeline	a
31	Which of the following specifies relationship amongst variables?	Scatter Plot	Line Chart	Area Chart	All of above	d

32	Which of the following specifies category Proportions?	Pie Chart	Histogram	Bar chart	All of above	d
33	Which of the following is category of timeline?	Variant Timeline	ER Timeline	Comparative Timeline	Modular Timeline	c
34	Information Visualization techniques are	Flow Chart	Time Line	DFD	All of above	d
35	Data visualization techniques are:	Flow Chart	Time Line	Pie Chart	None of these	c
36	Data visualization is related with...	Pictorial representations	numerical representation	numerical calculations	None of these	a
37	Which of the following follows interactive visualization approach?	Zoom+Pan	Focus+Context	Overview+Details	all of above	d
38	Which of the following are Use of data visualization	See context of data	Clear data understanding	finding pattern in data	all of above	d
39	Which of the following specifies relationship amongst variables?	Pie Chart	Histogram	Area Chart	None of these	c
40	Which of the following specifies category Proportions?	Pie Chart	Scatter Plot	Line Chart	None of these	a

UNIT SIX	SUB : 410243 DA					
Sr. No.	Questions	a	b	c	d	Ans
1	Precies and steady format data is_____	Structured Data	Un Structured Data	semi Structured Data	Quasi Structured Data	a
2	Inconsistant Data is_____	Structured Data	Un Structured Data	semi Structured Data	Quasi Structured Data	b
3	Format that self defines itself is_____	Structured Data	Un Structured Data	semi Structured Data	Quasi Structured Data	c
4	A little Bit inconsistant data is_____	Structured Data	Un Structured Data	semi Structured Data	Quasi Structured Data	d
5	XML is an example of_____	Structured Data	Un Structured Data	semi Structured Data	Quasi Structured Data	
6	RDBMS Follows_____	Structured Data	Un Structured Data	semi Structured Data	Quasi Structured Data	a
7	Watson is developed by_____	IBM	Microsoft	AT&T	Google	a
8	Hadoop is _____ based Framework.	C++	Python	JAVA	C#	c
9	Which of the following are components of Hadoop?	MAPREDUCE	YARN	HDFS	All of Above	d

10	Which of the following are components of HIVE?	JDBC	Thrift Server	CLI	All of Above	d
11	Mahout provides_____	JAVA Executable Libraries	C# Executables	Mountable Image Format	All of Above	a
12	Which of the following are components of HIVE?	FLATTEN	Thrift Server	Muster	None of these	b
13	Which of the following are components of HIVE?	FLATTEN	Thrift Server	Muster	All of above	b
14	Which of the following is components of Hadoop?	Fork	YARN	CLI	Metadata	b
15	RDBMS Follows_____	Structured Data	Un Structured Data	semi Structured Data	Quasi Structured Data	a
16	Which of the following is a clustering techique?	Fuzzy K means	Canopy	K-Means	All of above	d
17	Which of the following is HBASE Data Model Terminology?	Row	Table	Column	All of Above	d
18	Which of the following is not a classification techique?	Logistic Regression	Random Forest	Recommender Algo	Naïve Bayes	c
19	Which of the following is a classification techique?	Logistic Regression	Random Forest	Naïve Bayes	All of Above	d
20	Which of the following is HBASE Data Model Terminology?	Column Family	Cell	Timestamp	All of Above	d

21	Which of the following is a clustering technique?	Logistic Regression	Random Forest	K-Means	Naïve Bayes	c
22	Which of the following is HBASE Data Model Terminology?	Identifier	Variant	Timestamp	None of the above	c
23	Which of the following is not a classification technique?	Logistic Regression	Random Forest	K-Means	Naïve Bayes	c
24	Which of the following are components of HIVE?	FLATTEN	Thrift Server	Muster	None of these	b
25	Which of the following is HBASE Data Model Terminology?	Identifier	Variant	Column Qualifier	None of the above	c
26	Mahout provides_____	JAVA Executable Libraries	C# Executables	Mountable Image Format	None of the above	a
27	Which of the following is not a clustering technique?	Logistic Regression	Canopy	K-Means	Fuzzy K means	a
28	Which of the following is a clustering technique?	Fuzzy K means	Canopy	K-Means	All of above	d

29	Point out the correct statement.	Hadoop do need specialized hardware to process the data	Hadoop 2.0 allows live stream processing of real-time data	In Hadoop programming framework output files are divided into lines or records	None of the above	b
30	What was Hadoop named after?	Creator Doug Cutting's favorite circus act	Cutting's high school rock band	The toy elephant of Cutting's son	A sound Cutting's laptop made during Hadoop development	c
31	_____ programming model used to develop Hadoop-based applications that can process massive amounts of data.	MapReduce	Mahout	Oozie	None of the above	a
32	Which of the following is not a classification techique?	Logistic Regression	Random Forest	K-Means	Naïve Bayes	c

33	Which of the following are components of HIVE?	FLATTEN	Thrift Server	Muster	All of above	b
34	Which of the following is components of Hadoop?	Fork	YARN	CLI	None of above	b
35	Hadoop is a framework that works with a variety of related tools. Common cohorts include _____	MapReduce, Hive and HBase	MapReduce, MySQL and Google Apps	MapReduce, Hummer and Iguana	All of above	a
36	NoSQL databases is used mainly for handling large volumes of _____ data.	Structured Data	Un Structured Data	semi Structured Data	Quasi Structured Data	b
37	Which of the following is not a phase of Data Analytics Life Cycle?	Communication	Recall	Data Preparation	Model Planning	b
38	Which of the following is a NoSQL Database Type?	SQL	Document databases	JSON	All of above	b

39	Which of the following is not a NoSQL database	SQL Server	MongoDB	Cassandra	None of the above	a
----	--	------------	---------	-----------	-------------------	---

1. Data Analysis is a process of?

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

[View Answer](#)

Ans : **D**

Explanation: Data Analysis is a process of inspecting, cleaning, transforming and modeling data with the goal of discovering useful information, suggesting conclusions and supporting decision-making.

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

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

[View Answer](#)

Ans : **B**

Explanation: Predictive Analytics is major data analysis approaches not Predictive Intelligence.

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

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

[View Answer](#)

Ans : **A**

Explanation: In data analysis, two main statistical methodologies are used Descriptive statistics and Inferential statistics.

4. 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

[View Answer](#)

Ans : **C**

Explanation: In descriptive statistics, data from the entire population or a sample is summarized with numerical descriptors.

5. Data Analysis is defined by the statistician?

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

[View Answer](#)

Ans : **D**

Explanation: Data Analysis is defined by the statistician John Tukey in 1961 as "Procedures for analyzing data."

6. 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

[View Answer](#)

Ans : **A**

Explanation: answering yes/no questions about the data (hypothesis testing)

7. 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

[View Answer](#)

Ans : A

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

8. 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**

[View Answer](#)

Ans : D

Explanation: The branch of statistics which deals with development of particular statistical methods is classified as applied statistics.

9. 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

[View Answer](#)

Ans : **C**

Explanation: modeling relationships within the data (E.g. regression analysis).

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

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

[View Answer](#)

Ans : **A**

Explanation: Text Data Mining is the process of deriving high-quality information from text.

SUB : 410243 DA

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

SUB : 410243 DA

10.	Value tells the trustworthiness of data in terms of quality and accuracy. a. TRUE b. FALSE
11.	GFS consists of a _____ Master and _____ Chunk Servers a. Single, Single b. Multiple, Single c. Single, Multiple d. Multiple, Multiple
12.	Files are divided into _____ sized Chunks. a. Static b. Dynamic c. Fixed d. Variable
13.	_____ is an open source framework for storing data and running application on clusters of commodity hardware. a. HDFS b. Hadoop c. MapReduce d. Cloud
14.	HDFS Stores how much data in each clusters that can be scaled at any time? a. 32 b. 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 a. TRUE b. FALSE
16.	Hortonworks was introduced by Cloudera and owned by Yahoo. a. TRUE b. FALSE
17.	Hadoop YARN is used for Cluster Resource Management in Hadoop Ecosystem. a. TRUE b. FALSE
18.	Google Introduced MapReduce Programming model in 2004. a. TRUE b. FALSE
19.	_____ phase sorts the data & _____ creates logical clusters. a. Reduce, YARN b. MAP, YARN c. REDUCE, MAP d. MAP, REDUCE

SUB : 410243 DA

20.	There is only one operation between Mapping and Reducing is it True or False... a. TRUE b. FALSE
21.	_____ is factors considered before Adopting Big Data Technology. a. Validation b. Verification c. Data d. Design
22.	_____ for improving supply chain management to optimize stock management, replenishment, and forecasting; a. Descriptive b. Diagnostic c. Predictive d. Prescriptive
23.	which among the following is not a Data mining and analytical applications? a. profile matching b. social network analysis c. facial recognition d. Filtering
24.	_____ as a result of data accessibility, data latency, data availability, or limits on bandwidth in relation to the size of inputs. a. Computation-restricted throttling b. Large data volumes c. Data throttling d. 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. a. Lowering costs b. Increasing revenues c. Increasing productivity d. Reducing risk
26.	Which storage subsystem can support massive data volumes of increasing size. a. Extensibility b. Fault tolerance c. Scalability d. High-speed I/O capacity
27.	_____ provides performance through distribution of data and fault tolerance through replication a. HDFS b. PIG c. HIVE d. HADOOP

SUB : 410243 DA

28.	<p>_____ is a programming model for writing applications that can process Big Data in parallel on multiple nodes.</p> <ul style="list-style-type: none"> a. HDFS b. MAP REDUCE c. HADOOP d. HIVE
29.	<p>_____ takes the grouped key-value paired data as input and runs a Reducer function on each one of them.</p> <ul style="list-style-type: none"> a. MAPPER b. REDUCER c. COMBINER d. PARTITIONER
30.	<p>_____ is a type of local Reducer that groups similar data from the map phase into identifiable sets.</p> <ul style="list-style-type: none"> a. MAPPER b. REDUCER c. COMBINER d. PARTITIONER
31.	<p>While Installing Hadoop how many xml files are edited and list them ?</p> <ul style="list-style-type: none"> i. core-site.xml ii. hdfs-site.xml iii. mapred.xml iv. yarn.xml
32.	<p>Write the code for core-site.xml ?</p> <pre> <?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 ></pre>
33.	<p>Write the code for hdfs-site.xml ?</p>

SUB : 410243 DA

Sr. 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 a. 2 Only b. 1 and 2 c. 1 and 3 d. 2 and 3
2.	Sentiment Analysis is an example of 1. Regression 2. Classification 3. Clustering 4 Reinforcement Learning a. 1, 2 and 4 b. 1 and 3 c. 1, 2 and 3 d. 1 and 2
3.	Can decision trees be used for performing clustering? a. True b. 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. a. 1, 3 and 4 b. 1, 2 and 3 c. 1, 2 and 4 d. 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?

SUB : 410243 DA

	<p>a. There were 28 data points in clustering analysis b. The best no. of clusters for the analyzed data points is 4 c. The proximity function used is Average-link clustering d. The above dendrogram interpretation is not possible for K-Means clustering analysis</p>
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? <p>1. 1 2. 2 3. 3 4. 4</p>
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 a. 1 and 2 b. 2 and 3 c. 2 and 4 d. 1, 2 and 4
11.	The discrete variables and continuous variables are two types of a. Open end classification b. Time series classification c. Qualitative classification d. Quantitative classification

SUB : 410243 DA

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

SUB : 410243 DA

	prior knowledge of the data, they are among the most popular algorithms for classifying text documents. a. ID3 b. Naïve Bayes classifiers c. CART d. None of these.
20.	High entropy means that the partitions in classification are a. Pure b. Not pure c. Useful d. Useless
21.	Which of the following statements about Naive Bayes is incorrect? a. Attributes are equally important. b. Attributes are statistically dependent of one another given the class value. c. Attributes are statistically independent of one another given the class value. d. 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. a. Max Entropy is 1 b. Max Entropy is 2 c. Max Entropy is 3 d. 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. a. k-nearest neighbor is same as k-means b. k-means clustering is a method of vector quantization c. k-means clustering aims to partition n observations into k clusters d. 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 3. Regression 4. None of These

SUB : 410243 DA

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

Sr. 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. a. True b. 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 ? a. Yes b. 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

SUB : 410243 DA

7.	<p>There are _____ major Classification of Collaborative Filtering Mechanisms</p> <ol style="list-style-type: none"> 1. 1 2. 2 3. 3 4. None of These
8.	<p>Movie Recommendation to peoples is an example of</p> <ol style="list-style-type: none"> 1. User Based Recommendation 2. Item Based Recommendation 3. Knowledge Based Recommendation 4. Content Based Recommendation
9.	<p>_____ recommenders rely on an explicitly defined set of recommendation rules.</p> <ol style="list-style-type: none"> 1. Constraint Based 2. Case Based 3. Content Based 4. User Based
10.	<p>Parallelized hybrid recommender systems operate independently of one another and produce separate recommendation lists.</p> <ol style="list-style-type: none"> 1. True 2. False
11.	<p>Association rules are sometimes referred to as</p> <ol style="list-style-type: none"> a. market basket analysis b. Itemset Filtering c. Frequent Itemset Analysis d. None of these.
12.	<p>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</p> <ol style="list-style-type: none"> a. 0.4 b. 0.5 c. 0.6 d. 0.7
13.	<p>Lift is defined as the measure of certainty or trustworthiness associated with each discovered rule.</p> <ol style="list-style-type: none"> a. TRUE b. FALSE
14.	<p>_____ is able to identify trustworthy rules, but it cannot tell whether a rule is coincidental.</p> <ol style="list-style-type: none"> a. Lift b. Confidence c. Support d. Leverage

SUB : 410243 DA

	recommend items based on similarity measures between users and/or items. The items recommended to a user are those preferred by similar users.
15.	<p>a. Collaborative Filtering System</p> <p>b. Content Based Recommendation</p> <p>c. Knowledge Based Recommendation</p> <p>d. Hybrid Approaches</p>
16.	Pure collaborative approaches take a matrix of given user-item ratings as the only input and typically produce output. Is it Pure Collaborative? <p>a. Yes</p> <p>b. No</p>
17.	With respect to the determination of the set of similar users, one common measure used in recommender systems is <p>a. Cosine Similarity Measure</p> <p>b. Pearson's correlation coefficient.</p> <p>c. Mean Squared Error Method</p> <p>d. None of these.</p>
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. <p>a. Item-Based Recommendation</p> <p>b. User-Based Recommendation</p> <p>c. Content-Based Recommendation</p> <p>d. None of these</p>
19.	Here are two very short texts to compare and find the cosine similarity measure? <p>I. Julie loves me more than Linda loves me</p> <p>II. Jane likes me more than Julie loves me</p> <p>a. 0.6</p> <p>b. 0.7</p> <p>c. 0.8</p> <p>d. 0.9</p>
20.	_____ is based on the availability of item descriptions and a profile that assigns importance to these characteristics. <p>a. Item-Based Recommendation</p> <p>b. User-Based Recommendation</p> <p>c. Content-Based Recommendation.</p> <p>d. None of these</p>
21.	Consider the features of a movie which are not relevant to a recommendation system. <p>a. The set of actors of the movie.</p> <p>b. The Director</p> <p>c. The Year in which the movie was made</p> <p>d. The Budget of the movie.</p>

SUB : 410243 DA

	A _____ has been implemented, for similarity based retrieval under nearest neighbors.
22.	<p>a. k-nearest-neighbor method (kNN)</p> <p>b. Conventional Neural Network (CNN)</p> <p>c. Bayes Theorem</p> <p>d. Naïve Bayes Classifier</p>
23.	Case-based recommenders focus on the retrieval of similar items on the basis of different types of similarity measures <p>a. TRUE</p> <p>b. FALSE</p>
24.	In _____ recommendation approaches, items are retrieved using similarity measures that describe to which extent item properties match some given user's requirements. <p>a. Item-Based</p> <p>b. Case-Based</p> <p>c. Content-Based</p> <p>d. User-Based</p>
25.	_____ are based on a sequenced order of techniques, in which each succeeding recommender only refines the recommendations of its predecessor. <p>a. Weighted Hybrids</p> <p>b. Mixed Hybrids</p> <p>c. Cascade Hybrids</p> <p>d. Switching Hybrids</p>
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 <p>a. Weighted Hybrids</p> <p>b. Mixed Hybrids</p> <p>c. Cascade Hybrids</p> <p>d. Switching Hybrids</p>

SUB : 410243 DA

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

Answer

A

MCQ No - 2

What are the main components of Big Data?

- (A) MapReduce
- (B) HDFS
- (C) YARN
- (D) All of these

Answer

D

MCQ No - 3

What are the different features of Big Data Analytics?

- (A) Open-Source
- (B) Scalability
- (C) Data Recovery
- (D) All the above

Answer

D

MCQ No - 4

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

Answer

A

MCQ No - 5

What are the four V's of Big Data?

- (A) Volume
- (B) Velocity
- (C) Variety
- (D) All the above

SUB : 410243 DA

Answer

D

All of the following accurately describe Hadoop, EXCEPT:

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

Answer

B

MCQ No - 7

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

- (A) Mapreduce
- (B) Drill
- (C) Oozie
- (D) None of the above

Answer

A

MCQ No - 8

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

Answer

C

MCQ No - 9

SUB : 410243 DA

Big data analysis does the following except

- (A) Collects data
- (B) Spreads data
- (C) Organizes data
- (D) Analyzes data

Answer

B

MCQ No - 10

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) All of the above

Answer

B

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

Answer

C

MCQ No - 12

The unit of data that flows through a Flume agent is

- (A) Log
- (B) Row
- (C) Event

SUB : 410243 DA

(D) Record

Answer

C

MCQ No - 13

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

- (A) Data Ingestion
- (B) Data Processing
- (C) Data dissemination
- (D) Data Storage

Answer

C

MCQ No - 14

Check below the best answer to "which industries employ the use of so-called "Big Data" in their day to day operations?

- (A) Weather forecasting
- (B) Marketing
- (C) Healthcare
- (D) All of the above

Answer

D

MCQ No - 15

There are almost as many bits of information in the digital universe as there are stars in the actual universe?

- (A) True
- (B) False

Answer

A

SUB : 410243 DA

MCQ No - 16

The word 'Big data' was coined by

- (A) Roger Mousalas
- (B) John Philips
- (C) Simon Woods
- (D) Martin Green

Answer

A

MCQ No - 17

The word 'Big Data' was coined in the year

- (A) 2000
- (B) 1970
- (C) 1998
- (D) 2005

Answer

C

MCQ No - 18

Concerning the Forms of Big Data, which one of these is odd?

- (A) Structured
- (B) Unstructured
- (C) Processed
- (D) Semi-Structured

Answer

C

MCQ No - 19

Big Data applications benefit the media and entertainment industry by

- (A) Predicting what the audience wants
- (B) Ad targeting

SUB : 410243 DA

- (C)** Scheduling optimization
- (D)** All of the above

Answer

D

MCQ No - 20

The feature of big data that refers to the quality of the stored data is

-
- (A)** Variety
 - (B)** Volume
 - (C)** Variability
 - (D)** Veracity

Answer

D



ZEAL EDUCATION SOCIETY'S
ZEAL COLLEGE OF ENGINEERING AND RESEARCH
NARHE | PUNE -41 | INDIA
DEPARTMENT OF COMPUTER ENGINEERING



Name of the Teacher: Ms. P. S. Patil

Class: BE
AY: 2020-21

Subject: Data Analytics
SEM: II

UNIT-1

1)	What is Big Data? a) Huge amount of data b) Small amount of data c) Huge File d) Big Storage
Ans:	a
Explanation:	It is Huge amount of data
2)	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
Explanation:	Big data management and data mining
3)	What are the main components of Big Data? a)MapReduce b)HDFS c)YARN d)All of these
Ans:	d
Explanation:	All of these
4)	The sources of Big Data are a)Stock Exchange b)Transport Data c) Banking Data d) All of the Above
Ans:	d
Explanation:	
5)	Big Data Characteristics are: a) Structured data b) Semi-structured data c) Quasi-structured data d) All of the above
Ans:	d
Explanation:	
6)	B1 tends to provide reports, dashboards, and queries on business



ZEAL EDUCATION SOCIETY'S
ZEAL COLLEGE OF ENGINEERING AND RESEARCH
NARHE | PUNE -41 | INDIA
DEPARTMENT OF COMPUTER ENGINEERING



	questions for the current period or in the past.
	a) True b) False
Ans:	a
Explanation:	
7)	Big data can come in multiple forms, including structured and non-structured data
	a) True b) False
Ans:	a
Explanation:	
8)	BI problems tend to require highly structured data organized
	a) Rows b) Columns c) Accurate Reporting d) All of the Above
Ans:	d
Explanation:	
9)	EDW achieves the objective of reporting and sometimes the creation of dashboards, perform analysis on unstructured data
	a) High-value data is hard to reach and leverage b) Data moves in batches from EDW to local analytical tools c) Data Science projects will remain isolated d) All of the Above
Ans:	d
Explanation:	
10)	Drivers of Big Data
	a) Medical information b) Photos and video footage uploaded to the World Wide Web c) data extracts d) Both a and b
Ans:	d
Explanation:	
11)	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
Explanation:	
12)	Select from option which is not the phase of data analytics



ZEAL EDUCATION SOCIETY'S
ZEAL COLLEGE OF ENGINEERING AND RESEARCH
NARHE | PUNE -41 | INDIA
DEPARTMENT OF COMPUTER ENGINEERING



	<ul style="list-style-type: none">a) model planningb) testingc) discoveryd) operationalize
Ans:	b
Explanation:	
13)	Which phase of data analytics require more time to complete
	<ul style="list-style-type: none">a) Data preparationb) model buildingc) communicate resultsd) Discovery
Ans:	a
Explanation:	
14)	What is analytic sandbox?
	<ul style="list-style-type: none">a) Toolb) Separate repositoryc) data cleaningd) Data conditioning
Ans:	b
Explanation:	
15)	The person which provides analytic techniques and modeling is called as.
	<ul style="list-style-type: none">a) Data Engineerb) Data scientistc) Business userd) Project manager
Ans:	b
Explanation:	
16)	What is task of Project manager?
	<ul style="list-style-type: none">a) analytic modellingb) Provide requirementc) ensure meeting objectivesd) creates DB environment
Ans:	c



ZEAL EDUCATION SOCIETY'S
ZEAL COLLEGE OF ENGINEERING AND RESEARCH
NARHE | PUNE -41 | INDIA
DEPARTMENT OF COMPUTER ENGINEERING



Explanation:	
17)	Identifying Key Stakeholders this task is performed in which phase? a) Data preparation b) model building c) Discovery d) communicate results
Ans:	c
Explanation:	
18)	ETL process is performed in which phase a) Discovery b) communicate results c) model planning d) Data preparation
Ans:	d
Explanation:	
19)	How much data Data science teams prefer for analysis? a) too little b) average c) more d) more than average
Ans:	c
Explanation:	
20)	select from option tool which is not used in model planning phase a) Data wrangler b) R c) SQL Analysis service d) SAS/ACESS
Ans:	c
Explanation:	



ZEAL EDUCATION SOCIETY'S
ZEAL COLLEGE OF ENGINEERING AND RESEARCH
NARHE | PUNE -41 | INDIA
DEPARTMENT OF COMPUTER ENGINEERING



21)	if reports and dashboards will be impacted and need to change this task is performed by. <ul style="list-style-type: none">a) Project sponsorb) BI Analystc) Data Engineerd) Project manager
Ans:	b
Explanation:	
22)	What is need of data analytic lifecycle. <ul style="list-style-type: none">a) Data cleaningb) To solve Big data problemsc) Data conditioningd) Data Exploration
Ans:	b
Explanation:	
23)	How many phases are there in data analytic lifecycle? <ul style="list-style-type: none">a) 4b) 5c) 6d) 7
Ans:	c
24)	The person with technical skills is called as? <ul style="list-style-type: none">a) Business userb) Data Engineerc) Data scientistd) Project sponsor
Ans:	b
25)	What is outcome of Model building phase? <ul style="list-style-type: none">a) Analytic resultsb) Quality datac) Datad) Potential resources
Ans:	a



ZEAL EDUCATION SOCIETY'S
ZEAL COLLEGE OF ENGINEERING AND RESEARCH
NARHE | PUNE -41 | INDIA
DEPARTMENT OF COMPUTER ENGINEERING



Pravin S.Patil

Subject Teacher



ZEAL EDUCATION SOCIETY'S
ZEAL COLLEGE OF ENGINEERING AND RESEARCH
NARHE | PUNE -41 | INDIA
DEPARTMENT OF COMPUTER ENGINEERING



Name of the Teacher: Ms. P. S. Patil

Class: BE
AY: 2020-21

Subject: Data Analytics
SEM: II

UNIT-1I

1)	1. A statement made about a population for testing purpose is called? a) Statistic b) Hypothesis c) Level of Significance d) Test-Statistic
Ans:	b
Explanation:	
2)	If the assumed hypothesis is tested for rejection considering it to be true is called? a) Null Hypothesis b) Statistical Hypothesis c) Simple Hypothesis d) Composite Hypothesis
Ans:	a
Explanation:	
3)	A statement whose validity is tested on the basis of a sample is called? a) Null Hypothesis b) Statistical Hypothesis c) Simple Hypothesis d) Composite Hypothesis
Ans:	b
Explanation:	
4)	A hypothesis which defines the population distribution is called? a) Null Hypothesis b) Statistical Hypothesis c) Simple Hypothesis d) Composite Hypothesis
Ans:	c
Explanation:	
5)	If the null hypothesis is false then which of the following is accepted? a) Null Hypothesis b) Positive Hypothesis c) Negative Hypothesis d) Alternative Hypothesis.
Ans:	d
Explanation:	



ZEAL EDUCATION SOCIETY'S
ZEAL COLLEGE OF ENGINEERING AND RESEARCH
NARHE | PUNE -41 | INDIA
DEPARTMENT OF COMPUTER ENGINEERING



6)	The rejection probability of Null Hypothesis when it is true is called as? a) Level of Confidence b) Level of Significance c) Level of Margin d) Level of Rejection
Ans:	b
Explanation:	
7)	The point where the Null Hypothesis gets rejected is called as? a) Significant Value b) Rejection Value c) Acceptance Value d) Critical Value
Ans:	d
Explanation:	
8)	If the Critical region is evenly distributed then the test is referred as? a) Two tailed b) One tailed c) Three tailed d) Zero tailed
Ans:	a
Explanation:	
9)	The type of test is defined by which of the following? a) Null Hypothesis b) Simple Hypothesis c) Alternative Hypothesis d) Composite Hypothesis
Ans:	c
Explanation:	
10)	Which of the following is defined as the rule or formula to test a Null Hypothesis? a) Test statistic b) Population statistic c) Variance statistic d) Null statistic
Ans:	a
Explanation:	
11)	Type 1 error occurs when? a) We reject H ₀ if it is True b) We reject H ₀ if it is False c) We accept H ₀ if it is True d) We accept H ₀ if it is False
Ans:	a
Explanation:	
12)	The probability of Type 1 error is referred as? a) 1- α



ZEAL EDUCATION SOCIETY'S
ZEAL COLLEGE OF ENGINEERING AND RESEARCH
NARHE | PUNE -41 | INDIA
DEPARTMENT OF COMPUTER ENGINEERING



	b) β c) α d) $1-\beta$
Ans:	c
Explanation:	
13)	Alternative Hypothesis is also called as? a) Composite hypothesis b) Research Hypothesis c) Simple Hypothesis d) Null Hypothesis
Ans:	b
Explanation:	
14)	Which of the following is required by K-means clustering? a) defined distance metric b) number of clusters c) initial guess as to cluster centroids d) all of the mentioned
Ans:	d
Explanation:	
15)	Point out the wrong statement. a) k-means clustering is a method of vector quantization b) k-means clustering aims to partition n observations into k clusters c) k-nearest neighbor is same as k-means d) none of the mentioned
Ans:	c
Explanation:	
16)	Hierarchical clustering should be primarily used for exploration. a) True b) False
Ans:	a
Explanation:	
17)	Which of the following function is used for k-means clustering? a) k-means b) k-mean c) heatmap d) none of the mentioned
Ans:	a
Explanation:	
18)	Which of the following clustering requires merging approach? a) Partitional b) Hierarchical c) Naive Bayes d) None of the mentioned
Ans:	b
Explanation:	



ZEAL EDUCATION SOCIETY'S
ZEAL COLLEGE OF ENGINEERING AND RESEARCH
NARHE | PUNE -41 | INDIA
DEPARTMENT OF COMPUTER ENGINEERING



19)	K-means is not deterministic and it also consists of number of iterations.
	a) True b) False
Ans:	a
20)	Depending on acceptance and rejection of null hypothesis there are 2 types of error produced
	a) Type 1 b) Type 2 c) None of these d) All of these
Ans:	d
21)	The power of a test can be defined as a possibility of ...
	a) Rejecting null hypothesis b) Accepting null hypothesis c) Increasing null hypothesis d) Decreasing null hypothesis
Ans:	a
22)	For a fixed significance level, a greater sample size is mandatory to discover a
	a) Minor difference in mean b) Major difference in mean c) Average difference in mean d) None of the above
Ans:	a
23)	ANOVA tests if any of the population means vary from other population means
	a) True b) False
Ans:	a
24)	Clustering is defined as group of same kind of objects which are gathered by use of
	a) Unsupervised method b) Supervised method c) Semi supervised method d) None of these
Ans:	a
25)	Following are the applications of Kmeans
	a) Image Processing b) Medical c) Customer Segmentation d) All of the above



ZEAL EDUCATION SOCIETY'S
ZEAL COLLEGE OF ENGINEERING AND RESEARCH
NARHE | PUNE -41 | INDIA
DEPARTMENT OF COMPUTER ENGINEERING



Ans:

d

Pravin S.Patil

Subject Teacher

Unit-I

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

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

[View Answer](#)

Ans : C

Explanation: data in Peta bytes i.e. 10^{15} byte size is called Big Data.

2. How many V's of Big Data

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

[View Answer](#)

Ans : D

Explanation: Big Data was defined by the “3Vs” but now there are “5Vs” of Big Data which are Volume, Velocity, Variety, Veracity, Value

3. Transaction data of the bank is?

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

[View Answer](#)

Ans : A

Explanation: Data which can be saved in tables are structured data like the transaction data of the bank.

4. In how many forms BigData could be found?

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

[View Answer](#)

Ans : B

Explanation: BigData could be found in three forms: Structured, Unstructured and Semi-structured.

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

[View Answer](#)

Ans : D

Explanation: All of the above are Benefits of Big Data Processing.

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

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

[View Answer](#)

Ans : D

Explanation: Apache Pytarch is incorrect Big Data Technologies.

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%

[View Answer](#)

Ans : C

Explanation: The overall percentage of the world's total data has been created just within the past two years is 90%.

8) Which of the following step is performed by data scientist after acquiring the data?

- a) Data Cleansing

- b) Data Integration
- c) Data Replication
- d) All of the mentioned

Ans: Data Cleansing

9) 3V's are not sufficient to describe big data.

- a) True

- b) False

Ans: True

10. Communicative and collaborative is one among the key skill sets and behavioral characteristics of a data scientist [True / False]?

a. True

b. False

Answer : a

11. ----- are the sources of Bigdata [select all that apply]

- I. Book
- II. Facebook
- III. Genome sequence
- IV. Video Surveillance

Ans:

12. BI analyses the past data and make future predictions True/False ?

a. True

b. False

Answer : b

12. In which phase of data analytics ETLT is performed?

Ans: Phase 2 Data preparation is done in this phase. An analytical sandbox is used in this to perform analytics for the entire duration of the project. While you explore, preprocess and condition data, modeling follows suit. To get the data into the sandbox, you will perform ETLT (extract, transform, load and transform).

A. Discovery

B. Model Planning

C. Model Building

D. **Data Preparation**

13. In which data analytics lifecycle phase is an analytic sandbox prepared?

Phase 2 — Data preparation: Phase 2 requires the presence of an analytic sandbox, in which the team can work with data and perform analytics for the duration of the project. The team needs to execute extract, load, and transform (ELT) or extract, transform and load (ETL) to get data into the sandbox.

A. **Data Preparation**

B. Model Planning

C. Model Building

D. Discovery

14. In which phase would the team expect to invest most of the project time?

A. **Data Preparation**

B. Model Planning

C. Model Building

D. Discovery

15. In which phase would the team expect to invest least time of the project time?

A. Data Preparation

B. Model Planning

C. **Model Building**

D. Discovery

16. from following tools which tool is used for Model building?

- a. Hadoop
- b. Octave
- c. OpenRefine
- d. All of Above

Ans B

17. from following tools which tool is used for Data preparation

- a. Alpine Miner b. Excel c. Matlab d.Weka

Ans . A

18. To determine if the project was completed on time and within budget, is the key role of _____

- A. Project Sponsor
- B. Project Manager**
- C. Data Engineer
- D. Data Scientist

19. How many Phases are there in Data Analytics Lifecycle?

- A. 3
- B. 6**
- C. 7
- D. Any

20. In data Analytics life cycle we can move back and refine the work done. True or False

- A. True**
- B. False

21. What are the key outputs from Analytics Projects?

- A. PPT
- B.report
- C. code
- D. All of above**

22. _____ provides subject matter expertise for analytical techniques, data modeling and applying valid analytical techniques to give business problems.

- A. Project Sponsor
- B. Project Manager
- C. Data Engineer
- D. Data Scientist**

Unit-II

1. A statement about a population developed for the purpose of testing is called:

- (a) Hypothesis
- (b) Hypothesis testing
- (c) Level of significance
- (d) Test-statistic

Answer : a

2. Any hypothesis which is tested for the purpose of rejection under the assumption that it is true is called:

- (a) Null hypothesis
- (b) Alternative hypothesis
- (c) Statistical hypothesis
- (d) Composite hypothesis

Answer : a

3. A statement that is accepted if the sample data provide sufficient evidence that the null hypothesis is false is called:

- (a) Simple hypothesis
- (b) Composite hypothesis
- (c) Statistical hypothesis
- (d) Alternative hypothesis

Answer : d

4. The alternative hypothesis is also called:

- (a) Null hypothesis
- (b) Statistical hypothesis
- (c) Research hypothesis
- (d) Simple hypothesis

Answer : c

5. The probability of rejecting the null hypothesis when it is true is called:

- (a) Level of confidence
- (b) Level of significance
- (c) Power of the test
- (d) Difficult to tell

Answer : b

6. If the critical region is located equally in both sides of the sampling distribution of test-statistic, the test is called:

- (a) One tailed
- (b) Two tailed
- (c) Right tailed
- (d) Left tailed

Answer : b

7. The choice of one-tailed test and two-tailed test depends upon:

- (a) Null hypothesis
- (b) Alternative hypothesis
- (c) None of these
- (d) Composite hypotheses

Answer : b

8. Test of hypothesis $H_0: \mu = 50$ against $H_1: \mu > 50$ leads to:

- (a) Left-tailed test
- (b) Right-tailed test
- (c) Two-tailed test
- (d) Difficult to tell

Answer : b

9. Testing $H_0: \mu = 25$ against $H_1: \mu \neq 25$ leads to:

- (a) Two-tailed test
- (b) Left-tailed test
- (c) Right-tailed test
- (d) Neither (a), (b) and (c)

Answer : a

10. A formula that provides a basis for testing a null hypothesis is called:

- (a) Test-statistic
- (b) Population statistic
- (c) Both of these
- (d) None of the above

Answer : a

11. $1 - \alpha$ is also called:

- (a) Confidence coefficient
- (b) Power of the test
- (c) Size of the test
- (d) Level of significance

Answer : a

12. Area of the rejection region depends on:

- (a) Size of α
- (b) Size of β
- (c) Test-statistic
- (d) Number of values

Answer : a

13. Student's t-test is applicable only when:

- (a) $n \leq 30$ and σ is known
- (b) $n > 30$ and σ is unknown
- (c) $n = 30$ and σ is known
- (d) All of the above

Answer : a

14. In an unpaired samples t-test with sample sizes $n_1 = 11$ and $n_2 = 11$, the value of tabulated t should be obtained for:

- (a) 10 degrees of freedom
- (b) 21 degrees of freedom
- (c) 22 degrees of freedom
- (d) 20 degrees of freedom

Answer : d

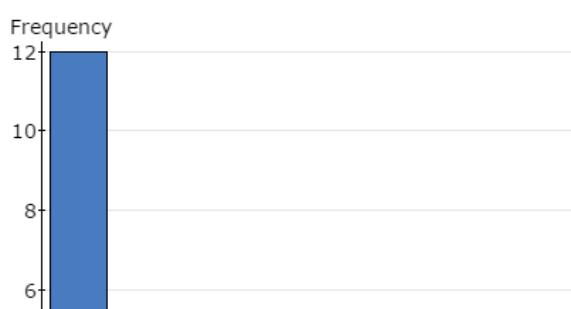
15. The purpose of statistical inference is:

- (a) To collect sample data and use them to formulate hypotheses about a population
- (b) To draw conclusion about populations and then collect sample data to support the conclusions
- (c) To draw conclusions about populations from sample data
- (d) To draw conclusions about the known value of population parameter

Answer : c

16. The histogram to the right represents the hospital length of stay (in days) for patients at a nearby medical facility. How many patients are included in the histogram?

- a. 5
- b. 21
- c. 17
- d. 9



Answer : b

17. Using the histogram to the right that represents the hospital lengths of stay (in days) for patients at a nearby medical facility, determine the relationship between the mean and the median.

- a. Mean = Median
- b. Mean \approx Median
- c. Mean < Median
- d. Mean > Median

Answer : d

18. The statement “If there is sufficient evidence to reject a null hypothesis at the 10% significance level, then there is sufficient evidence to reject it at the 5% significance level” :

Please select the best answer of those provided below.

- a. Always True
- b. Never True
- c. Sometimes True; the p-value for the statistical test needs to be provided for a conclusion
- d. Not Enough Information; this would depend on the type of statistical test used

Answer : c

19. Analysis of variance in short form is?

- a) ANOV
- b) AVA
- c) ANOVA
- d) ANVA

Ans:c

20) Which of the following is required by K-means clustering?

- a) defined distance metric
- b) number of clusters
- c) initial guess as to cluster centroids
- d) all of the mentioned

Ans: defined distance metric, number of clusters, initial guess as to cluster centroids

21) Hierarchical clustering should be primarily used for exploration.

- a) True
- b) False

Ans: True

22) Which of the following function is used for k-means clustering?

- a) k-means
- b) k-mean
- c) heatmap
- d) none of the mentioned

Ans: k-means

23) The goal of clustering a set of data is to

- a)divide them into groups of data that are near each other
- b)choose the best data from the set
- c)determine the nearest neighbors of each of the data
- d)predict the class of data

Ans: divide them into groups of data that are near each other

24) The k-means algorithm...

- a)always converges to a clustering that minimizes the mean-square vector-representative distance
- b)can converge to different final clustering, depending on initial choice of representatives
- c)is widely used in practice
- d)is typically done by hand, using paper and pencil
- e)should only be attempted by trained professionals

Ans: can converge to different final clustering, depending on initial choice of representatives, is widely used in practice

25) Considering the K-means algorithm, after current iteration, we have 3 centroids $(0, 1)$ $(2, 1)$, $(-1, 2)$. Will points $(2, 3)$ and $(2, 0.5)$ be assigned to the same cluster in the next iteration?

- a) Yes

- b) No

Ans: Yes

26) What are the two types of Hierarchical Clustering?

- a)Top-Down Clustering (Divisive)
- b)Bottom-Top Clustering (Agglomerative)
- c)Dendrogram

d)K-means

Ans: Top-Down Clustering (Divisive), Bottom-Top Clustering (Agglomerative)

27) The most commonly used measure of similarity is the _____ or its square.

- a)euclidean distance
- b)city-block distance
- c)Chebychev's distance
- d)Manhattan distance

Ans: euclidean distance

29) Which of the following is required by K-means clustering?

- a)defined distance metric
- b)number of clusters
- c)initial guess as to cluster centroids

Ans: defined distance metric, number of clusters, initial guess as to cluster centroids

30) Clustering is a-

- A. Supervised learning
- B. Unsupervised learning
- C. Reinforcement learning
- D. None

Ans: Unsupervised learning

31) Which of the following clustering algorithms suffers from the problem of convergence at local optima?

- A. K- Means clustering
- B. Hierarchical clustering
- C. Diverse clustering
- D. All of the above

Ans: K- Means clustering, Hierarchical clustering, Diverse clustering

32) Which version of the clustering algorithm is most sensitive to outliers?

- A. K-means clustering algorithm
- B. K-modes clustering algorithm
- C. K-medians clustering algorithm
- D. None

Ans: K-means clustering algorithm

33) Which of the following is a bad characteristic of a dataset for clustering analysis-

- A. Data points with outliers
- B. Data points with different densities
- C. Data points with non-convex shapes
- D. All of the above

Ans: Data points with outliers, Data points with different densities, Data points with non-convex shapes

34) For clustering, we do not require-

- A. Labeled data
- B. Unlabeled data
- C. Numerical data
- D. Categorical data

Ans: Labeled Data

35) Which of the following is an application of clustering?

- A. Biological network analysis
- B. Market trend prediction
- C. Topic modeling
- D. All of the above

Ans: Biological network analysis, Market trend prediction, Topic modeling

36) The final output of Hierarchical clustering is-

- A. The number of cluster centroids
- B. The tree representing how close the data points are to each other
- C. A map defining the similar data points into individual groups
- D. All of the above

Ans: The tree representing how close the data points are to each other

37. Which type of test is the Wilcoxon rank sum test?

- a. Parametric
- b. non parametric
- c. Distributed
- d. Normal

38. Input data for Wilcoxon test is normally distributed, True or False?

39. What is the null hypothesis for a Wilcoxon test?

- a. Two group means are equal.
- b. Two or more group means are equal.
- c. Two mean groups are not equal.
- d. None of these

40 Which of following test statics is used in Wilcoxon Rank Sum Test?

- a. test statistics \leq critical value, H_0 will be Rejected
- b. if test statistics $>$ critical value, H_0 will be Rejected
- c. if test statistics $>$ critical value, H_0 will be accepted
- d. none of these.

40. What must you include when applying Wilcoxon Rank sum test?

- a. variance
- b. Critical Value
- c. Rank sum
- e. standard deviation

Type1 and type 2 error

41. Type 1 is also called as

- a. False Positive
- b. false negative
- c. True Positive
- d. True negative

42. Type 2 is also called as

- a. False Positive
- b. False negative
- c. True Positive
- d. True negative

43. Type 1 error occurs when_____

- a. Null hypothesis rejected when it is true.
- b. Null hypothesis is accepted when it is false
- c. Null hypothesis rejected when it is false

d. None of Above

44. Type 2 error occurs when _____

- a. Null hypothesis rejected when it is true.
- b. Null hypothesis is accepted when it is false
- c. Null hypothesis rejected when it is false
- d. All of above

ANOVA

44. Analysis of Variance is statistical method of comparing_____ of several populations.

- a. Means
- b. variance
- c. standard Deviation
- d. None of above.

45. ANOVA is used when_____

- a. If more than two population
- b. for two population
- c. for Three population
- d. for any populations

46. What is Null Hypothesis in ANOVA?

- a. all group means are equal
- b. Three group means are equal
- c. atleast one pair of group means unequal.
- d. all group means are unequal.

47. What do ANOVA calculate?

a. Z-score

b. F ratio

c. T-score

d. Chi Square

Q.25 What are the two types of variance which can occur in your data?

a. Independent and Dependent

b. Between and within groups

c. Personal and interpersonal

d. Anova and Anoca

Q.26 If between group mean sum of square variability increases value of F statistics _____

a. Increases

b. Decreases

c. Neutral

d. None of these

Q.27 What must you include when applying ANOVA test?

a. Means

b. Critical Value

c. degree of freedom

d. F statistics

e. All of above

Q.28 How many dependent variables are there in a two-way ANOVA?

a.1

b.3

c.2

d.any

Q.29 Which of following test statics is used in ANOVA?

a.if critical value > F ratio, H_0 will be Rejected

b.if critical value < F ratio, H_0 will be Rejected

c.if critical value > F ratio, H_0 will be accepted

d.None of these

Q.30 Various types of ANOVA are____.

a.Two way ANOVA

b.ANCova

c.MANOVA

d.ZANOVA

Unit-III

1.A collection of one or more items is called as _____

- (A)Itemset
- (B)Support
- (C)Confidence
- (D)Support Count

Ans:A

2.Frequency of occurrence of an itemset is called as _____

- (A)Support
- (B)Confidence
- (C)Support Count
- (D)Rules

Ans:C

3.An itemset whose support is greater than or equal to a minimum support threshold is _____

- (A)Itemset
- (B)Frequent Itemset
- (C)Infrequent items
- (D)Threshold values

Ans:B

4.What does FP growth algorithm do?

- (A)It mines all frequent patterns through pruning rules with lesser support
- (B)It mines all frequent patterns through pruning rules with higher support
- (C)It mines all frequent patterns by constructing a FP tree
- (D)It mines all frequent patterns by constructing an itemsets

Ans:C

5.What techniques can be used to improve the efficiency of apriori algorithm?

(A)Hash-based techniques

(B)Transaction Increases

(C)Sampling

(D)Cleaning

Ans:A

6. Linear Regression is a supervised machine learning algorithm.

A) TRUE

B) FALSE

Ans:A

7. It is possible to design a Linear regression algorithm using a neural network?

A) TRUE

B) FALSE

Ans:A

8. Which of the following methods do we use to find the best fit line for data in Linear Regression?

A) Least Square Error

B) Maximum Likelihood

C) Logarithmic Loss

D) Both A and B

Ans:A

9. A local retailer has a database that stores 10,000 transactions of last summer. After analyzing the data, a data science team has identified the following statistics:
• {battery} appears in 6,000 transactions.
• {sunscreen} appears in 5,000 transactions.
• {sandals} appears in 4,000 transactions.
• {bowls} appears in 2,000 transactions.
• {battery, sunscreen} appears in 1,500 transactions.
• {battery, sandals} appears in 1,000 transactions.
• {battery, bowls} appears in 250 transactions.
• {battery, sunscreen, sandals} appears in 600 transactions.
Q) What are the confidence values of {battery} -> {sunscreen} and {battery, sunscreen} -> {sandals} ?

- a) 0.3 and 0.4
- b) 0.25 and 0.4
- c) 0.25 and 0.15
- d) 0.6 and 0.4

Ans: b

10. Which of the following implies no relationship with respect to correlation?

- a) $\text{Cor}(X, Y) = 1$
- b) $\text{Cor}(X, Y) = 0$
- c) $\text{Cor}(X, Y) = 2$
- d) All of the mentioned

Ans:b

11. If Linear regression model perfectly fit i.e., train error is zero, then

- a) Test error is also always zero
- b) Test error is non zero
- c) Couldn't comment on Test error
- d) Test error is equal to Train error

Ans:C

12. Which of the following metrics can be used for evaluating regression models?

- i) R Squared
- ii) Adjusted R Squared
- iii) F Statistics
- iv) RMSE / MSE / MAE

- a) ii and iv
- b) i and ii
- c) ii, iii and iv
- d) i, ii, iii and iv

Ans:d

13.How many coefficients do you need to estimate in a simple linear regression model (One independent variable)?

- a) 1
- b) 2
- c) 3
- d) 4

Ans:b

14.In a simple linear regression model (One independent variable), If we change the input variable by 1 unit. How much output variable will change?

- a) by 1
- b) no change
- c) by intercept
- d) by its slope

Ans:d

15.Function used for linear regression in R is _____

- a) lm(formula, data)
- b) lr(formula, data)
- c) lrm(formula, data)
- d) regression.linear(formula, data)

Ans:a

16. In syntax of linear model lm(formula,data,..), data refers to _____

- a) Matrix
- b) Vector
- c) Array
- d) List

Ans:b

17. In the mathematical Equation of Linear Regression $Y = \beta_1 + \beta_2X + \epsilon$, (β_1, β_2) refers to _____

- a) (X-intercept, Slope)
- b) (Slope, X-Intercept)
- c) (Y-Intercept, Slope)
- d) (slope, Y-Intercept)

Ans:c

18. _____ is an incredibly powerful tool for analyzing data.

- a) Linear regression
- b) Logistic regression
- c) Gradient Descent
- d) Greedy algorithms

Ans:a

19. The square of the correlation coefficient r^2 will always be positive and is called the _____

- a) Regression
- b) Coefficient of determination

- c) KNN
- d) Algorithm

Ans:b

20. Predicting y for a value of x that's outside the range of values we actually saw for x in the original data is called _____

- a) Regression
- b) Extrapolation
- c) Intrapolation
- d) Polation

Ans:b

21. What is predicting y for a value of x that is within the interval of points that we saw in the original data called?

- a) Regression
- b) Extrapolation
- c) Intrapolation
- d) Polation

Ans:c

22. _____ is a simple approach to supervised learning. It assumes that the dependence of Y on X_1, X_2, \dots, X_p is linear.

- a) Linear regression
- b) Logistic regression
- c) Gradient Descent
- d) Greedy algorithms

Ans:a

23. Although it may seem overly simplistic, _____ is extremely useful both conceptually and practically.

- a) Linear regression
- b) Logistic regression
- c) Gradient Descent
- d) Greedy algorithms

Ans:a

24. _____ refers to a group of techniques for fitting and studying the straight-line relationship between two variables.

- a) Linear regression
- b) Logistic regression
- c) Gradient Descent
- d) Greedy algorithms

Ans:a

25. What do you mean by support(A)?

- a. Total number of transactions containing A
- b. Total Number of transactions not containing A
- c. Number of transactions containing A / Total number of transactions
- d. Number of transactions not containing A / Total number of transactions

Ans: c

Data Processing and Analysis

Unit 4

Multiple Choice Questions with Answer Key

1. What is a hypothesis?

- a. A statement that the researcher wants to test through the data collected in a study.
- b. A research question the results will answer.
- c. A theory that underpins the study.
- d. A statistical method for calculating the extent to which the results could have happened by chance.

Answer: a

2. Qualitative data analysis is still a relatively new and rapidly developing branch of research methodology.

- a. True
- b. False

Answer: a

3.. The process of marking segments of data with symbols, descriptive words, or category names is known as _____.

- a. Concurring
- b. Coding
- c. Colouring
- d. Segmenting

Answer: b

4. What is the cyclical process of collecting and analysing data during a single research study called?

- a. Interim analysis
- b. Inter analysis
- c. Inter-item analysis
- d. Constant analysis

Answer: a

5. The process of quantifying data is referred to as _____.

- a. Typology
- b. Diagramming
- c. Enumeration
- d. Coding

Answer: c

6. An advantage of using computer programs for qualitative data is that they _____.

- a. Can reduce time required to analyse data (i.e., after the data are transcribed)
- b. Help in storing and organising data
- c. Make many procedures available that are rarely done by hand due to time constraints
- d. All of the above

Answer: d

7. Boolean operators are words that are used to create logical combinations.

- a. True
- b. False

Answer: a

8. _____ are the basic building blocks of qualitative data.

- a. Categories
- b. Units
- c. Individuals
- d. None of the above

Answer: a

9. This is the process of transforming qualitative research data from written interviews or field notes into typed text.

- a. Segmenting
- b. Coding
- c. Transcription
- d. Mnemoning

Answer: c

10. A challenge of qualitative data analysis is that it often includes data that are unwieldy and complex; it is a major challenge to make sense of the large pool of data.

- a. True
- b. False

Answer: a

11. Hypothesis testing and estimation are both types of descriptive statistics.

- a. True
- b. False

Answer: b

12. A set of data organised in a participants(rows)-by-variables(columns) format is known as a “data set.”

- a. True
- b. False

Answer: a

13. A graph that uses vertical bars to represent data is called a ____

- a. Line graph
- b. Bar graph
- c. Scatterplot
- d. Vertical graph

Answer: b

14. _____ are used when you want to visually examine the relationship between two quantitative variables.

- a. Bar graphs
- b. Pie graphs
- c. Line graphs
- d. Scatterplots

Answer: d

15. The denominator (bottom) of the z-score formula is

- a. The standard deviation
- b. The difference between a score and the mean
- c. The range
- d. The mean

Answer: a

16. Which of these distributions is used for a testing hypothesis?

- a. Normal Distribution
- b. Chi-Squared Distribution
- c. Gamma Distribution
- d. Poisson Distribution

Answer b

17. A statement made about a population for testing purpose is called?

- a. Statistic
- b. Hypothesis
- c. Level of Significance
- d. Test-Statistic

Answer: b

18. If the assumed hypothesis is tested for rejection considering it to be true is called?

- a. Null Hypothesis
- b. Statistical Hypothesis
- c. Simple Hypothesis
- d. Composite Hypothesis

Answer: a

19. If the null hypothesis is false then which of the following is accepted?

- a. Null Hypothesis
- b. Positive Hypothesis
- c. Negative Hypothesis
- d. Alternative Hypothesis.

Answer: d

20. Alternative Hypothesis is also called as?

- a. Composite hypothesis
- b. Research Hypothesis
- c. Simple Hypothesis
- d. Null Hypothesis

Answer: b

	marks	question	A	B	C	D	ans
0	1	A group of 4 bits is also called?	Nibble	Byte	Kb	None	4 bits make one nibble.
1	1	There are how many types of Big Data:	3	2	1	None	Big Data is of 3 types.
2	1	Which of the following are the V's of Big Data:	All	Volume	Variety	Velocity.	This is an explanation.
3	1	Which of these is not a characteristic of Big data?	Storage	Volume	Variety	Velocity.	This is an explanation.
4	2	Which of the following is a drawback of Big Data:	Cost	Significant	Process	Fraud Detection	Big Data requires high cost to maintain huge amount of data
5	2	Fullform of GINA is:	Global Innovation Network and Analysis.	Global Invention in Networks and Analytics	Globally Investment in Neurons and Analytics	None	GINA stands for Global Innovations Networks and Analysis.
6	2	Which is the phase 3 in Data Analytics Life cycle.	Model Planning	Model Building	Data Preparation	Operationalize	Model Planning is the 3rd phase in life cycle.
7	2	GINA team thought to accomplish mainly goals:	3	2	1	5	GINA targeted to achieve three goals for the project.
8	2	The Data Preparation stage doesn't involve:	Analyzation	Collection	Cleansing	Processing.	This is an explanation.
9	2	Unstructured Data is further divided into how many types?	2	3	4	5	Unstructured data is divided into 2 types.
10	2	The GINA team mainly used which software tool to analyze the Data	Tableau	Hadoop	HIVE	SQL	The team used Tableau to visualize the Data.
11	2	Which of the following is the first step of Data Analytics Life Cycle:	Discovery	Data Preparation.	Model Planning	Data Aware	This is an explanation.
12	2	There are how many phases in data analytics life cycle:	6	5	4	7	there are 6 stages in data analytics life cycle.
13	2	SEMMA Methodology has how many stages:	5	4	6	7	SEMMA methodology has five stages.
14	2	Which phase of Life Cycle requires collaboration with stakeholders?	Phase 5	Phase 6	Phase 4	Phase 3	Phase 5 involves collaboration with stakeholders.
15	2	In Building a Model, how many phases are required:	2	3	4	5	This is an Explanation.
16	2	How much Data in the whole world is structured:	0.2	0.4	0.6	0.5	Only 20% of world's total data is structured.
17	2	10^7 bytes of memory is equal to:	1ZB	1TB	1YB	1XB	10^7 B is equal to 1 ZB.
18	2	Data Scientists in the GINA team used which technique on the textual Description of the Innovation Roadmap Idea.	Natural Language Processing(NLP)	Hadoop	HIVE	SQL	NLP technique was used on the description of Innovation Roadmap Idea.
19	2	How many types of data analytics methodologies are there?	2	4	3	6	Two types of data analytical methodologies are there. EDA and CDA
20	3	Other name for Bell Curve is:	Normal Distribution.	Poisson Distribution	Binomial Distribution	Bernoulli Distribution.	Bell Curve is also known as normal distribution.
21	3	One of the most important tasks in big data analytics is:	Statistical Modeling	Testing of Data	Visualization	Operationalize	One of the most important tasks in big data analytics is statistical modeling
22	3	Some of the approaches considered for building the data analytics lifecycle framework best practices are:	All	CRISP-DM	SEMMA	MAD Skills	This is an explanation.
23	3	In Phase 4, the team develops datasets for:	All	Testing of Data	Training of Data	Production purposes	This is an explanation.
24	3	Fullform of CRISP-DM Methodology is:	Cross Industry Standard Process for Data Mining	Cross International Standard Process for Data Modeling	Common Industry Standard Program for Data Mining	Company's Initial Standards Progress for Data Methods	CRISP-DM stands for Cross Industry Standard Process for Data Mining.
25	3	SEMMA Methodology doesn't include which of the following stages:	Evaluate	Sample	Explore	Asses	This is an Explanation.
26	3	In Which stage, the data is monitored and analyzed to see if the generated model is creating the expected results.	Operationalize	Collection	Plan Model	Data Aware	In last phase i.e. Operationalize Data is monitored and analyzed to see if the generated model is creating the expected results.
27	3	Data is captured in how many ways:	3	4	5	6	Data is captured in 3 main ways.

	marks	question	A	B	C	D	ans
28	3	In phase 2 of the Data Analytics Life Cycle, the team performs how many analytics to get the data in the sandbox.	3	2	4	6	The team performs ETL and ELT and ETLT in 2nd phase of the cycle.
29	3	The total area under the bell curve is _____ unit.	1	2	3	4	Area under the bell curve is 1 unit.
30	1	Wilcoxon rank-sum test is also known as?	Mann-Whitney U test	Mean Difference	Alternative Hypothesis	Null Hypothesis	Wilcoxon rank-sum test is also called Mann-Whitney U Test.
31	1	Which test is also known as T-test?	Hypothesis Test	Mean Difference	K-means test	None	This is an explanation.
32	1	This equation is of which test?	Mean Difference	K-Means	Null Hypothesis	Alternative Hypothesis	This eqn is of Mean difference test.
33	1	A test of a statistical hypothesis, where the region of rejection is on a side of the sampling distribution, is called _____.	One tailed test	Two-tailed test	Tailed test	Null test	A test of a statistical hypothesis, where the region of rejection is on only one side of the sampling distribution, is called a one-tailed test
34	1	How many types of Statistical Hypothesis is there?	2	3	4	6	There are two types of Statistical Hypothesis.
35	1	Analysis of Variance is also referred as?	ANOVA	Mean Difference	Alternative Hypothesis	Null Hypothesis	ANOVA stands for Analysis of Variance.
36	1	How many steps are involved in a Hypothesis Testing?	4	2	3	5	There are 4 steps in Hypothesis testing.
37	2	The strength of evidence in support of a null hypothesis is measured by?	P-value	K-value	H-value	Null-value	The strength of evidence in support of a null hypothesis is measured by the P-value.
38	2	Difference in means is also called?	Two sample t-test	T- test	M-test	Two sample test	Difference in means is also known as two sample t test.
39	2	The k-medoids is also called _____ algorithm.	Partitioning Around Medoids (PAM)	Lloyd's Algorithm	Poisson's Algorithm	Regression	The k-medoids is also called partitioning around medoids (PAM) algorithm .
40	2	Clustering is an example of _____?	Unsupervised Learning	Supervised Learning	Classification	Regression	Clustering is an example of unsupervised learning.
41	2	Which of the following is not an advantage of K means Clustering?	Requires a Priori	Fast	Robust	easy to evaluate.	This is an explanation.
42	2	The probability of committing a Type 2 error is called	Beta	Alpha	Delta	Theta	The probability of committing a Type II error is called Beta
43	2	The _____ variation we have within clusters, the more homogeneous (similar) the data points are within the same cluster.	Less	More	Variable	Fixed	The less variation we have within clusters, the more homogeneous (similar) the data points are within the same cluster.
44	2	Which hypothesis is usually the hypothesis in which sample observations result is purely from chance?	Null-Hypothesis	Mean Difference	K-means test	Alternative Hypothesis	Null Hypothesis is usually the hypothesis that sample observations result purely from chance.
45	2	Classical" ANOVA for balanced data does how many things at once?	3	2	1	4	Classical" ANOVA for balanced data does three things at once.
46	2	K-mean clustering is used to solve which problems?	NP-hard problems	NP Problems	Hypothesis Problems	P problems	NP hard problems are solved using K means clustering.
47	2	The probability of committing a Type I error is called?	Alpha	Beta	Gama	Delta	The probability of committing a Type I error is called alpha
48	2	K means Clustering is also known as?	Lloyd's Algorithm	Gaussian Algorithm	Poisson's Algorithm	None	K means clustering is also called Lloyds algo.
49	3	Which algorithm requires the user to specify the number of clusters k to be generated.	K-means clustering	Gaussian Algorithm	Alternative Hypothesis	Null Hypothesis	k-means clustering requires the user to specify the number of clusters k to be generated.
50	3	K means clustering uses which approach to solve the problems?	Expectation-maximization	Greedy Approach	Divide and Conquer	None	expectation-maximization technique is used by k means clustering.
51	3	How many factors affect the power of a hypothesis test?	3	2	1	4	The power of a hypothesis test is affected by three factors.
52	3	Law of Variance is called?	Eve's Law	Laplace Law	Poisson's Algorithm	Regression	Law of variance is also called Eve's law.
53	3	K-Medoids use which approach to solve problems?	Greedy Approach	Divide and Conquer	Recursive	None	K Medoids use greedy approach to solve problems
54	3	The time complexity of k means clustering is?	O(n^2)	O(nlogn)	O(n)	O(1)	Time complexity is O(n^2) of k means clustering.
55	3	the number (k) of clusters assumed in k-medoids is known as?	Priori	Null Hypothesis	ANNOVA	Effect size	The number k of clusters assumed known as priori.

	marks	question	A	B	C	D	ans
56	3	What is the difference between the true value and the value specified in the null hypothesis.	Effect -size	Null Hypothesis	Alternative Hypothesis	ANOVA	The effect size is the difference between the true value and the value specified in the null hypothesis.
57	3	Time complexity of k medoids is?	$O(n^2)$	$O(n \log n)$	$O(n)$	$O(n^3)$	This is an explanation.
58	3	Which algorithm aims at minimizing an objective function know as squared error function	K-means	Mean Difference	Alternative Hypothesis	ANOVA	K means algorithm aims at minimizing an objective function know as squared error function
59	1	Which algorithm was the earliest of the association rule algorithms? n	Apriori Algorithm	Gaussian Algorithm	K means clustering	Bernoulli Distribution.	Apriori Algorithm was earliest in the association of algorithms.
60	1	The Apriori algorithm takes a _____ iterative approach to uncovering the frequent itemsets by first determining all the possible items	Bottom-Up	Top-Down	Recursive	None	The Apriori algorithm takes a bottom-up iterative approach to uncovering the frequent itemsets by first determining all the possible items
61	1	Apriori uses which structure to count candidate item sets efficiently?	BFS	DFS	Queue	Stack	Apriori uses breadth-first search and a Hash tree structure to count candidate item sets efficiently
62	1	"y=a+b*x^2". This equation shows which regression?	Polynomial Regression	Logistic Regression	Linear Regression	Lasso Regression	This is an explanation.
63	2	_____ is defined as the measure of certainty or trustworthiness associated with each discovered rule.	Confidence	Recursion	Item-set	None	Confidence is defined as the measure of certainty or trustworthiness associated with each discovered rule.
64	2	In which Regression, we predict the value by 1 or 0?	Logistic Regression	Linear Regression	Both	None	In Logistic Regression, we predict the value by 1 or 0.
65	2	The formula for linear regression is:	$Y' = bX + A$	$Y' = bX - A$.	$Y' = bX / A$.	$Y' = bX * A$.	The formula for linear regression is: $Y' = bX + A$.
66	2	Which regression is useful when there are a large number of independent variables.	Partial Least Squares(PLS) Regression	Cox Regression	Lasso Regression	Logistic Regression	PLS regression is also useful when there are a large number of independent variables.
67	2	Which regression is an approach for predicting a response using a single feature.	Linear-Regression	Logistic Regression	Elasticnet Regression	None	Simple linear regression is an approach for predicting a response using a single feature.
68	2	Association rule mining consists of _____ steps.	2	3	4	5	Association rule mining consists of 2 steps
69	2	Which type of regression is suitable when dependent variable is ordinal in nature?	Ordinal Regression	Linear Regression	Cox Regression	Logistic Regression	Ordinal regression is suitable when dependent variable is ordinal in nature
70	2	Which regression is used for support vector machines	ElasticNet Regression	Linear Regression	Logistic Regression	None	ElasticNet regression is used for support vector machines,
71	2	Which regression can solve both linear and non-linear models?	Support Vector Regression	Linear Regression	Logistic Regression	ElasticNet Regression	Support-Vector Regression can solve both linear and non linear models.
72	2	Which is the most common method used for fitting a regression line	Least Square Method	Mean Difference	Null Hypothesis	Classification	Least Square Method is the most common method used for fitting a regression line
73	2	_____ problems are when the output variable is a real or continuous value.	Regression	Classification	Recursive	Hypothesis	A regression problem is when the output variable is a real or continuous value.
74	2	Linear Regression is a machine learning algorithm based on _____ learning regression model.	Supervised Learning	Unsupervised Learning	Recursive Learning	All	Linear Regression is a machine learning algorithm based on supervised regression algorithm.
75	2	When dependent variable's variability is not equal across values of an independent variable, it is called	Heteroscedasticity	Homooscedasticity	Multicollinearity	Outliers.	When dependent variable's variability is not equal across values of an independent variable, it is called heteroscedasticity
76	2	requires large sample sizes because maximum likelihood estimates are less powerful at low sample sizes than ordinary least square	Logistic Regression	Linear Regression	Lasso Regression	ElasticNet Regression	Logistic Regression requires large sample sizes because maximum likelihood estimates are less powerful at low sample sizes than ordinary least square
77	2	PCR Regression is divided into how many steps?	2	3	4	5	PCR regression is divided into 2 steps
78	3	L2 regularization is also called?	Tikhonov Regularization	Norm Regularization	Poisson's Regularization	None	This is an explanation.
79	3	When the variance of count data is greater than the mean count, it is a case of?	Overdispersion	Underdispersion	Dispersion	High dispersion	When the variance of count data is greater than the mean count, it is a case of overdispersion

	marks	question	A	B	C	D	ans
80	3	Which regression assumes the normal distribution of the dependent variable?	Linear-Regression	Logistic Regression	Elasticnet Regression	None	Linear regression assumes the normal or gaussian distribution of the dependent variable.
81	3	Nature of predicted data in regression is?	Ordered	Unordered	Both	None	Nature of predicted data in regression is ordered.
82	3	Which regression uses a binary dependent variable but ignores the timing of events.	Logistic Regression	Linear Regression	Cox Regression	Lasso Regression	Logistic regression uses a binary dependent variable but ignores the timing of events.
83	3	The Ridge Regression is also known as?	Shrinkage Regression	Percentile Regression	Elasticnet Regression	Lasso Regression	The ridge regression is also known as Shrinkage Regression.
84	3	In which regression, we calculate Root Mean Square Error(RMSE) to predict the next weight value.	Linear-Regression	ElasticNet Regression	Logistic Regression	All	In Linear Regession we calculate Root Mean Square Error(RMSE) to predict the next weight value.
85	3	The _____ is the standard deviation of the observed residuals.	Residual standard error	Mean Difference Error	Data Error	All	The residual standard error is the standard deviation of the\observed residuals.
86	3	Which Regression is used when dependent variable has count data.	Poisson Regression	Linear Regression	Cox Regression	Lasso Regression	Poisson regression is used when dependent variable has count data.
87	3	_____ regression can handle both over-dispersion and under-dispersion.\n	Quasi-Poisson regression	Cox Regression	Elasticnet Regression	Linear Regression	Quasi-Poisson regression can handle both over-dispersion and under-dispersion.\n
88	3	_____ is the regularization parameter in Lasso Regression?	λ	θ	Ω	β	λ is the regularization parameter in lasso regression.
89	1	Decision Tree is a hierarchical model that does the separation of the\input space into class regions using:	Recursion	Pointers	Greedy Approach	Divide and Conquer	Decision Tree is a hierarchical model that recursively does the separation of the\input space into class regions
90	1	Learning Algorithm of Decision Tree is:	Greedy Approach	Divide and Conquer	Both	None	Decision Tree uses greedy approach for learning algorithm.
91	1	Normal Distribution is also called?	Gaussian Distribution	Bernoulli Distribution	Naïve Bias	Binary Distribution	This is an explanation.
92	1	Classification has how many phases:	2	3	4	5	There are 2 phases of classification.
93	1	"Every pair of features being classified is independent of each other".This principle is used by:	Naïve Bias Classifier	Decision Tree	Bernoulli Distribution	Normal Distribution	Naïve Bias uses the principle that every pair of features being classified is independent of each other.
94	2	This equation is of which theorem?	Gaussian Distribution	Binary Distribution	Naïve Bias	Gross-Entrpoy	This is an explaination.
95	2	In Naïve Bias, The Datasets are divided into how many types?	2	3	4	5	data sets are divided into two types in naïve bias.
96	2	Decision trees can be used to predict non-categorical values is called?	Regression Trees	Categorial trees	Normal tree	None	Decision trees can be used to predict non-categorical values is called regression trees
97	2	An attribute with____ Gini index should be preferred in a decision tree.	Lower	Higher	Recursive	Negative	an attribute with lower Gini index should be preferred.
98	2	In Naïve Bias, if any two events A and B are independent, then,	$P(A,B)=P(A)P(B)$	$P(A,B)=P(A)/P(B)$	$P(A,B)=P(B)/P(A)$	$P(A,B)=P(B)P(A)$	If any two events A and B are independent, then, $P(A,B)=P(A)P(B)$
99	2	What is the measure of uncertainty of a random variable in a decision tree.	Entropy.	Gain	Gini Index	None	Entropy is the measure of uncertainty of a random variable
100	2	Which of the following is not true for decision trees?	Stable	Easy to understand	Easy to explain	Easy to evaluate.	this is an explaination.
101	2	Decision tree algorithm falls under the category of which learning?	Supervised	Unsupervised	Regression	Classification	Decision tree algorithm falls under the category of supervised learning
102	2	False Positives and False Negatives is an application of which theorem?	Bayes' Theorem	Binary Distribution	Bernoulli Distribution	Normal Distribution	One of the use Bayes Theorem is false positives and false negatives.
103	2	Decision Tree used in mining the data are of how many types?	2	3	4	5	There are 2 types of decision trees used in data mining.
104	3	In Bayes' Theorem, $P(A)$ and $P(B)$ are the probabilities of observing A and B respectively; they are known as:	Marginal Probability	Normal Distribution	Bernoulli Distribution	Parallel Algorithm.	$P(A)$ and $P(B)$ are the probabilities of observing A and B respectively; they are known as the marginal probability.

	marks	question	A	B	C	D	ans
105	3	ID3 Algorithm in a decision tree stands for?	Iterative Dichotomiser 3 (ID3)	Interval Driven	Interconnected Decision	None	ID3 stands for Iterative Dichotomiser 3 (ID3)
106	3	Probably the best way of estimating performance for very small data sets is:	Boot Strapped Method	Normal Distribution	Naïve Bias	Binary Distribution	Probably the best way of estimating performance for very small data sets is bootstrapped method
107	3	The Decision Tree works on which form?	Disjunctive Normal Form	Product of Sum	Bijective Form	Conjunctive Form	Decision Tree works on Disjunctive normal form.
108	3	The decoupling of the class conditional feature distributions means that each distribution can be independently estimated as a _____ distribution.	1-D	2-D	3-D	NONE	The decoupling of the class conditional feature distributions means that each distribution can be independently estimated as a one dimensional distribution.
109	3	Theoretical concept to evaluate Classifiers is:	COLT	PAC Model	Naïve Bias	Prediction.	This is an explanation.
110	3	_____ is a metric to measure how often a randomly chosen element would be incorrectly identified	Gini Index	Entropy	Pointer	Gross-Entrpoy	Gini Index is a metric to measure how often a randomly chosen element would be incorrectly identified
111	3	The most notable types of decision tree algorithms are:	3	2	1	4	The most notable types of decision tree algorithms are 3
112	3	Which process is completed when the subset at a node all has the same value of the target variable?	Recursive Partitioning	Termination	Transformation	Prediction.	The recursive partition is completed when the subset at a node all has the same value of the target variable
113	3	The _____ method reserves a certain amount for testing and uses the remainder for training.	Holdout	Parallel Algorithm	Naïve Bias	Normal Distribution	The holdout method reserves a certain amount for testing and uses the remainder for training
114	3	This equation is of which theorem?	Bayes' Theorem	Normal Distribution	Bernoulli Distribution	Gross-Entrpoy	This is an explanation.
115	3	"Independence among the features". This is an assumption in:	Naïve Bais Classifier	Bernoulli Distribution	Parallel Algorithm	Binary Distribution	Independence among the features is an assumption in Naïve bias.
116	3	Error rate obtained from training data is called:	Resubstitution Error	Grid	Gini Index	True error	error rate obtained from training data is called resubstitution error.
117	3	In Decision Tree entropy is _____ to content.	proportional	inverse	High	Less	This is an explanation.
118	3	In Decision Tree, No root-to-leaf path should contain the same discrete attribute _____.	Twice	Once	Thrice	Four Times.	No root-to-leaf path should contain the same discrete attribute twice
119	1	Using _____, designers can make information understandable for stakeholders.	Data Visualization	Classification	Regression	Supervised Learning.	Using data visualization methods, designers can make information understandable for stakeholders.
120	1	The additional visual methods include:	All	Tree Map	Parallel Coordinates	Semantic Networks.	This is an explanation.
121	1	Data Visualization tools Doesn't include:	Ms--Excel	Tableau	Power BI	Jupyter	This is an explanation.
122	1	Which of the following requires Javascript Knowledge to run the visualization tool?	All	Chart.js	Polymap	Sigmajs	This is an explanation.
123	1	Merits of Tableau doesn't include which factor:	Cost	Performance	Usage	Computation	Merits of tableau doesn't include the cost factor.
124	1	Which of these is not a type of Big Data Visualization.	Pictograph	Bar-Graph	Line-Chart	Pie-Chart	This is an explanation.
125	2	The drag-and-drop editor od which tool makes it easy to create professional-looking designs without a lot of visual design skill.	Infogram	Google Chart	Tableau	Grafana	The drag-and-drop editor of Infogram makes it easy to create professional-looking designs without a lot of visual design skill.
126	2	How many V's are defined for Data Visualization.	4	6	2	3	There are 4 V's of Data visualization.
127	2	Which of the following is not a free Data Visualization tool?	Tableau	Google Chart	Jupyter	Hub-Spot CRM	Tableau is a chargeable tool of data visualization.
128	2	Companies that work with both traditional and big data use which technique to look at customer segments or market shares?	Pie-Chart	Bar-Graph	Stream graph	Line-Chart	Companies that work with both traditional and big data may use pie chart to look at customer segments or market shares
129	2	Visualization of Data includes which of the following problems:	All	Information Loss	Visual Noise	Large Image Perception.	This is an explanation.
130	2	Mainly, Data Visualization has how many types of challenges?	5	6	4	2	There are 5 main challenges to data visualization.

	marks	question	A	B	C	D	ans
131	2	Which tool uses HTML5/SVG to visualize data	Google Charts	Jupyter	Grafana	Tableau	Google charts uses HTML5/SVG since its browser compatible.
132	2	According to Colin Ware's Information Visualization: Perception for Design, he defines _____ pre-attentive visual properties.	4	2	1	3	According to Colin Ware's Information Visualization: Perception for Design, he defines four pre-attentive visual properties
133	2	_____ is based on space-filling visualization of hierarchical data.	Tree-Map	Stream graph	Bar-graph	Line-Chart	Tree map method is based on space-filling visualization of hierarchical data
134	2	Which graph shows the dependency relationships between activities and current schedule status.	Gantt-Chart	Line-Chart	Pie-Chart	Bar-Graph	Gant chart show the dependency relationships between activities and current schedule status.
135	2	Another name for distribution free data is:	Non parametric data	Parametric Data	static data	Dynamic data	Non parametric data is also called distribution free data.
136	2	Which chart is used for comparison of values, such as sales performance for several persons or businesses in a single time.	Bar-Graph	Gantt-Graph	Line-Chart	Pie-Chart	Bar Graph is used for Comparison of values, such as sales performance for several persons or businesses in a single time
137	2	_____ are graphics in the field of statistics used to visualize quantitative data.	Graphical-Techniques	Line-Chart	Regression	Classification	Graphical Techniques are graphics in the field of statistics used to visualize quantitative data.
138	2	_____ can handle several factors for a large number of objects per single screen, so it satisfies the data variety criterion.	Parallel Coordinates	Stream graph	Google Chart	Jupyter	Parallel Coordinates can handle several factors for a large number of objects per single screen, so it satisfies the data variety criterion
139	3	Chart.js provides how many types of charts?	8	5	3	6	This is an explanation.
140	3	Which visualization tool supports mixed data sources, annotations, and customizable alert functions, and it can be extended via hundreds of available plugins.	Grafana	Tableau	Google Chart	Jupyter	Grafana supports mixed data sources, annotations, and customizable alert functions, and it can be extended via hundreds of available plugins.
141	3	Which tool was created specifically for adding charts and maps to news stories.	Data Wrapper	Tableau	Google Chart	Jupyter	Datawrapper was created specifically for adding charts and maps to news stories.
142	3	Conventional Visualization methods doesn't include:	Mekko Chart	Pie-Chart	Bar-graph	Histogram	Mekko chart is a new technique to visualize data.
143	3	_____ is a type of a stacked area graph, which is displaced around a central axis, resulting in flowing and organic shape.	Streamgraph	Bar-Graph	Pie-Chart	Line-Chart	Streamgraph is a type of a stacked area graph, which is displaced around a central axis, resulting in flowing and organic shape
144	3	Which visual tool includes over 150 chart types and 1,000 map types?	Fusion charts	Tableau	Google Chart	Jupyter	Fusion charts includes over 150 chart types and 1,000 map types
145	3	Which graph/chart is a graphical representation of logical relationship between different concepts. It generates directed graph, the combination of nodes or vertices, edges or arcs, and label over each edge.	Semantic Networks	Bar-Graph	Pie-Chart	Line-Chart	A semantic network is a graphical representation of logical relationship between different concepts. It generates directed graph, the combination of nodes or vertices, edges or arcs, and label over each edge
146	3	According to SAS we can process only _____ of information per second on a flat screen.	1 Kilobit	1 Byte	1 Bit	1 MB	According to SAS we can process only 1 kilobit of information per second on a flat screen
147	3	There are _____ steps for interactive data visualization:	4	5	3	6	This is an explanation.
148	3	When working with big data, companies can use which visualization technique to track total application clicks by weeks, the average number of complaints to the call center by months, etc.\n\n	Line-Chart	Bar-Graph	Pie-Chart	Stream graph	When working with big data, companies can use the line chart visualization technique to track total application clicks by weeks, the average number of complaints to the call center by months, etc.\n\n
149	1	Which of the following Enterprises use HBase?	All	Facebook	Netflix	Adobe	This is an explanation.

	marks	question	A	B	C	D	ans
150	1	Which NLP is used in the present era?	Neural NLP	Symbolic NLP	Statistical NLP	None	From 2010, Neural NLP is being used.
151	1	The Computer World magazine states that unstructured information might account for more than _____ of all data in organizations.	70-80%	0.9	0.5	0.6	The Computer World magazine states that unstructured information might account for more than 70%-80% of all data in organizations.
152	1	Almost all of the information we use and share every day, such as articles, documents and e-mails, are completely _____.	Unstructured	Structured	Semantic	None	Almost all of the information we use and share every day, such as articles, documents and e-mails, are completely or partly unstructured
153	1	Which standard provided a common framework for processing information to extract meaning and create structured data about the information?	Unstructured Information Management Architecture (UIMA)	Management Architecture for Data	Data Architecture	None	The Unstructured Information Management Architecture (UIMA) standard provided a common framework for processing this information to extract meaning and create structured data about the information.
154	2	The base Apache Hadoop framework is composed of the how many modules?	4	2	3	6	The base Apache Hadoop framework is composed of the four modules.
155	2	No-SQL doesn't include which software?	MS-SQL	HBASE	DyanoDB	MongoDB	This is an explanation.
156	2	There are _____ main types of OLAP systems.	3	2	5	6	There are 3 types of OLAP systems.
157	2	SQL alternative in Apache HIVE is called?	HIVEQL	BASEQL	SPARK-QL	H-QL	HIVE-QL is the alternative to SQL in Apache Hive family.
158	2	MapReduce program executes in how many stages?	3	2	5	4	MapReduce program executes in three stages.
159	2	How many types of NO-SQL database are there?	4	3	2	6	There are 4 types of databases in NO-SQL.
160	2	MapReduce is a processing technique and a program model for distributed computing based on which programming Language?	JAVA	Python	C++	R	MapReduce is a processing technique and a program model for distributed computing based on java
161	2	Hive supports how many properties of transactions?	4	3	2	1	Hive supports all four properties of transactions
162	2	HDFS consists of only one Name Node that is called as?	Master Node	Slave Node	Both	None	HDFS consists of only one Name Node that is called the Master Node.
163	2	Which Apache Software is needed to process massive amounts of data for the purposes of natural-language search?	Apache HBASE	Apache Spark	Apache-PIG	Apache-mahout	Hbase to process massive amounts of data for the purposes of natural-language search
164	2	Which database store data in a format other than relational tables	NO-SQL	HIVESQL	SPARK-QL	H-QL	No-sql databases that store data in a format other than relational tables.
165	2	Which is a project of the Apache Software Foundation to produce free implementations of distributed or otherwise scalable machine learning algorithms focused primarily on linear algebra?	Apache Mahout	Apache Spark	Apache-PIG	Apache HBASE	Mahout is a project of the Apache Software Foundation to produce free implementations of distributed or otherwise scalable machine learning algorithms focused primarily on linear algebra.
166	2	Which model is a specialization of the split-apply-combine strategy for data analysis?	MapReduce	Hadoop	HBASE	HIVE	MapReduce model is a specialization of the split-apply-combine strategy for data analysis.
167	2	All Hadoop commands are invoked by which command?	\$SHADOOP_HOME/bin/hadoop	\$SHADOOP/bin/hadoop	\$HADOOP_HOME/hadoop	\$SHADOOP_HOME/bin	All Hadoop commands are invoked by the \$SHADOOP_HOME/bin/hadoop command
168	3	The table typically enforces the schema when the data is loaded into the table. This enables the database to make sure that the data entered follows the representation of the table as specified by the table definition. This design is called?	Schema on Write	Schema on Read	Schema for Read Write	None	The table typically enforces the schema when the data is loaded into the table. This enables the database to make sure that the data entered follows the representation of the table as specified by the table definition. This design is called schema on write.

	marks	question	A	B	C	D	ans
169	3	Which command formats the DFS filesystem?	Namenode -format	Node -format	Name -format	Format	Namenode -format command formats the DFS file system.
170	3	Which command applies the offline fsimage viewer to an fsimage?	oiv	fs	fc	ov	oiv applies the offline fsimage viewer to an fsimage.
171	3	Hadoop requires which Java Runtime Environment (JRE) or higher version?	1.6	1.2	1.5	1	Hadoop requires Java Runtime Environment (JRE) 1.6 or higher
172	3	Every Data node sends a Heartbeat message to the Name node every _____ seconds and conveys that it is alive.	3	2	4	1	Every Data node sends a Heartbeat message to the Name node every 3 seconds and conveys that it is alive
173	3	HDFS can store files upto:	1 TB	1 GB	1ZB	1PB	HDFS can store upto 1 TB of files.
174	3	Which of the following is a wide-column store?	HBase	SQL	DyanoDB	MongoDB	HBASE is a popular wide column store.
175	3	Which node acts as both a DataNode and TaskTracker in Hadoop.	Slave Node	Data Node	Admin Node	Name Node	A slave or worker node acts as both a DataNode and TaskTracker.
176	3	HDFS system uses which protocol for communication?	TCP/IP	TCP	UDP	IP	HDFS system uses TCP/IP sockets for communication
177	3	HDFS has how many services?	5	4	2	6	HDFS has five services.
178	3	_____ is a data warehouse software project built on top of Apache Hadoop for providing data query and analysis	Apache HIVE	Apache Spark	Apache-PIG	Apache HBASE	HIVE is a data warehouse software project built on top of Apache Hadoop for providing data query and analysis

Hadoop Online Quiz

Following quiz provides Multiple Choice Questions (MCQs) related to **Hadoop Framework**. You will have to read all the given answers and click over the correct answer. If you are not sure about the answer then you can check the answer using **Show Answer** button. You can use **Next Quiz** button to check new set of questions in the quiz.



Q 1 - HDFS block size is larger as compared to the size of the disk blocks so that

- A - Only HDFS files can be stored in the disk used.
- B - The seek time is maximum
- C - Transfer of a large files made of multiple disk blocks is not possible.
- D** - A single file larger than the disk size can be stored across many disks in the cluster.

Answer : D

[Hide Answer](#)

Q 2 - Zookeeper ensures that

- A - All the namenodes are actively serving the client requests
- B** - Only one namenode is actively serving the client requests

C - A failover is triggered when any of the datanode fails.

D - A failover can not be started by hadoop administrator.

Answer : B

[Hide Answer](#)

Q 3 - For a HDFS directory the replication factor(RF) is

A - same as the RF of the files in that directory

B - Zero

C - 3

D - Does not apply.

Answer : D

[Show Answer](#)

Q 4 - The hadfs command put is used to

A - Copy files from local file system to HDFS.

B - Copy files or directories from local file system to HDFS.

C - Copy files from from HDFS to local filesystem.

D - Copy files or directories from HDFS to local filesystem.

Answer : B

[Hide Answer](#)

Q 5 - Which of the below apache system deals with ingesting streaming data to

hadoop

- A - Ozie
- B - Kafka
- C - Flume
- D - Hive

Answer : C[Hide Answer](#)**Q 6 - YARN stands for**

- A - Yahoo's another resource name
- B - Yet another resource negotiator
- C - Yahoo's archived Resource names
- D - Yet another resource need.

Answer : B[Hide Answer](#)**Q 7 - When you increase the number of files stored in HDFS, The memory required by namenode**

- A - Increases
- B - Decreases
- C - Remains unchanged
- D - May increase or decrease

Answer : A

[Hide Answer](#)

Q 8 - In a HDFS system with block size 64MB we store a file which is less than 64MB. Which of the following is true?

- A - The file will consume 64MB
- B - The file will consume more than 64MB
- C** - The file will consume less than 64MB.
- D - Can not be predicted.

Answer : C

[Hide Answer](#)

Q 9 - What is distributed cache?

- A - The distributed cache is special component on name node that will cache frequently used data for faster client response. It is used during reduce step.
- B** - The distributed cache is special component on data node that will cache frequently used data for faster client response. It is used during map step.
- C - The distributed cache is a component that caches java objects.
- D - The distributed cache is a component that allows developers to deploy jars for Map-Reduce processing.

Answer : B

[Hide Answer](#)

Q 10 - In order to apply a combiner, what is one property that has to be satisfied by the values emitted from the mapper?

- A - Combiner can be applied always to any data
- B - Output of the mapper and output of the combiner has to be same key value pair and they can be heterogeneous
- C** - Output of the mapper and output of the combiner has to be same key value pair. Only if the values satisfy associative and commutative property it can be done.

Answer : C

[Hide Answer](#)

New Quiz

Hadoop Online Quiz

Following quiz provides Multiple Choice Questions (MCQs) related to **Hadoop Framework**. You will have to read all the given answers and click over the correct answer. If you are not sure about the answer then you can check the answer using **Show Answer** button. You can use **Next Quiz** button to check new set of questions in the quiz.



Q 1 - What is the main problem faced while reading and writing data in parallel from multiple disks?

- A - Processing high volume of data faster.
- B - Combining data from multiple disks.
- C - The software required to do this task is extremely costly.
- D - The hardware required to do this task is extremely costly.

Answer : B

[Hide Answer](#)

Q 2 - Under HDFS federation

- A - Each namenode manages metadata of the entire filesystem.
- B - Each namenode manages metadata of a portion of the filesystem.

C - Failure of one namenode causes loss of some metadata availability from the entire filesystem.

D - Each datanode registers with each namenode.

Answer : B

[Hide Answer](#)

Q 3 - For a HDFS directory the replication factor(RF) is

A - same as the RF of the files in that directory

B - Zero

C - 3

D - Does not apply.

Answer : D

[Hide Answer](#)

Q 4 - When the namenode finds that some blocks are over replicated, it

A - Stops the replication job in the entire hdfs file system.

B - It slows down the replication process for those blocks

C - It deletes the extra blocks.

D - It leaves the extra blocks as it is.

Answer : C

[Hide Answer](#)

Q 5 - The nature of hardware for the namenode should be

- A - Superior than commodity grade
- B - Commodity grade
- C - Does not matter
- D - Just have more Ram than each of the data nodes

Answer : A[Show Answer](#)**Q 6 - When a machine is declared as a datanode, the disk space in it**

- A - Can be used only for HDFS storage
- B - Can be used for both HDFS and non-HDFS storage
- C - Cannot be accessed by non-hadoop commands
- D - cannot store text files.

Answer : B[Show Answer](#)**Q 7 - When you increase the number of files stored in HDFS, The memory required by namenode**

- A - Increases
- B - Decreases
- C - Remains unchanged
- D - May increase or decrease

Answer : A

[Hide Answer](#)

Q 8 - A running job in hadoop can

- A - Be killed with a command
- B - Can never be killed with a command
- C - Can be killed only by shutting down the name node
- D - Be paused and run again

Answer : A

[Hide Answer](#)

Q 9 - What is writable?

- A - Writable is a java interface that needs to be implemented for streaming data to remote servers.
- B - Writable is a java interface that needs to be implemented for HDFS writes.
- C - Writable is a java interface that needs to be implemented for MapReduce processing.
- D - None of these answers are correct.

Answer : C

[Hide Answer](#)

Q 10 - In order to apply a combiner, what is one property that has to be satisfied by the values emitted from the mapper?

- A - Combiner can be applied always to any data
- B - Output of the mapper and output of the combiner has to be same key value pair and they can be heterogeneous
- C - Output of the mapper and output of the combiner has to be same key value pair. Only if the values satisfy associative and commutative property it can be done.

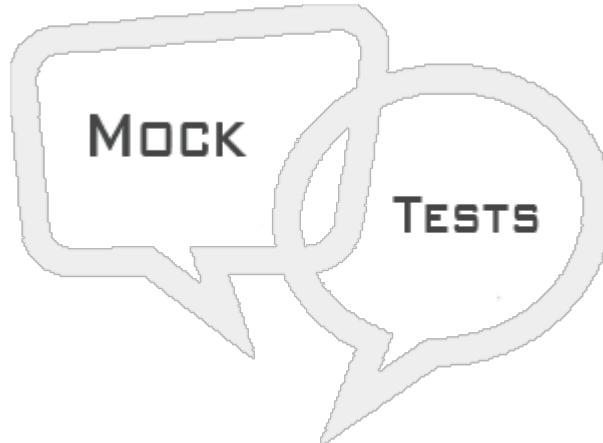
Answer : C

[Hide Answer](#)

New Quiz

HADOOP MOCK TEST

This section presents you various set of Mock Tests related to **Hadoop Framework**. You can download these sample mock tests at your local machine and solve offline at your convenience. Every mock test is supplied with a mock test key to let you verify the final score and grade yourself.



HADOOP MOCK TEST I

Q 1 - The concept using multiple machines to process data stored in distributed system is not new.

The High-performance computing HPC uses many computing machines to process large volume of data stored in a storage area network SAN. As compared to HPC, Hadoop

- A - Can process a larger volume of data.
- B - Can run on a larger number of machines than HPC cluster.
- C - Can process data faster under the same network bandwidth as compared to HPC.
- D - Cannot run compute intensive jobs.

Q 2 - Hadoop differs from volunteer computing in

- A - Volunteers donating CPU time and not network bandwidth.
- B - Volunteers donating network bandwidth and not CPU time.
- C - Hadoop cannot search for large prime numbers.
- D - Only Hadoop can use mapreduce.

Q 3 - As compared to RDBMS, Hadoop

- A - Has higher data Integrity.
- B - Does ACID transactions
- C - IS suitable for read and write many times
- D - Works better on unstructured and semi-structured data.

Q 4 - What is the main problem faced while reading and writing data in parallel from multiple disks?

- A - Processing high volume of data faster.
- B - Combining data from multiple disks.**
- C - The software required to do this task is extremely costly.
- D - The hardware required to do this task is extremely costly.

Q 5 - Which of the following is true for disk drives over a period of time?

- A - Data Seek time is improving faster than data transfer rate.
- B - Data Seek time is improving more slowly than data transfer rate.**
- C - Data Seek time and data transfer rate are both increasing proportionately.
- D - Only the storage capacity is increasing without increase in data transfer rate.

Q 6 - Data locality feature in Hadoop means

- A - store the same data across multiple nodes.
- B - relocate the data from one node to another.
- C - co-locate the data with the computing nodes.**
- D - Distribute the data across multiple nodes.

Q 7 - Which of these provides a Stream processing system used in Hadoop ecosystem?

- A - Solr
- B - Tez
- C - Spark**
- D - Hive

Q 8 - HDFS files are designed for

- A - Multiple writers and modifications at arbitrary offsets.
- B - Only append at the end of file**
- C - Writing into a file only once.
- D - Low latency data access.

Q 9 - A file in HDFS that is smaller than a single block size

- A - Cannot be stored in HDFS.
- B - Occupies the full block's size.
- C - Occupies only the size it needs and not the full block.**

D - Can span over multiple blocks.

Q 10 - HDFS block size is larger as compared to the size of the disk blocks so that

A - Only HDFS files can be stored in the disk used.

B - The seek time is maximum

C - Transfer of a large files made of multiple disk blocks is not possible.

D - A single file larger than the disk size can be stored across many disks in the cluster.

Q 11 - In a Hadoop cluster, what is true for a HDFS block that is no longer available due to disk corruption or machine failure?

A - It is lost for ever

B - It can be replicated from its alternative locations to other live machines.

C - The namenode allows new client request to keep trying to read it.

D - The Mapreduce job process runs ignoring the block and the data stored in it.

Q 12 - Which utility is used for checking the health of a HDFS file system?

A - fchk

B - fsck

C - fsch

D - fcks

Q 13 - Which command lists the blocks that make up each file in the filesystem.

A - hdfs fsck / -files -blocks

B - hdfs fsck / -blocks -files

C - hdfs fchk / -blocks -files

D - hdfs fchk / -files -blocks

Q 14 - The datanode and namenode are respectively

A - Master and worker nodes

B - Worker and Master nodes

C - Both are worker nodes

D - None

Q 15 - In the local disk of the namenode the files which are stored persistently are –

A - namespace image and edit log

B - block locations and namespace image

C - edit log and block locations

D - Namespace image, edit log and block locations.

Q 16 - When a client communicates with the HDFS file system, it needs to communicate with

A - only the namenode

B - only the data node

C - both the namenode and datanode

D - None of these

Q 17 - What mechanisms Hadoop uses to make namenode resilient to failure.

A - Take backup of filesystem metadata to a local disk and a remote NFS mount.

B - Store the filesystem metadata in cloud.

C - Use a machine with at least 12 CPUs

D - Using expensive and reliable hardware.

Q 18 - The main role of the secondary namenode is to

A - Copy the filesystem metadata from primary namenode.

B - Copy the filesystem metadata from NFS stored by primary namenode

C - Monitor if the primary namenode is up and running.

D - Periodically merge the namespace image with the edit log.

Q 19 - For the frequently accessed HDFS files the blocks are cached in

A - the memory of the datanode

B - in the memory of the namenode

C - Both A&B

D - In the memory of the client application which requested the access to these files.

Q 20 - User applications can instruct the namenode to cache the files by

A - adding cache file names to cache pool

B - adding cache config to cache pool

C - adding cache directive to cache pool

D - passing the file names as parameters to the cache pool

Q 21 - In Hadoop 2.x release HDFS federation means

- A - Allowing namenodes to communicate with each other.
- B - Allow a cluster to scale by adding more datanodes under one namenode.
- C - Allow a cluster to scale by adding more namenodes.**
- D - Adding more physical memory to both namenode and datanode.

Q 22 - Under HDFS federation

- A - Each namenode manages metadata of the entire filesystem.
- B - Each namenode manages metadata of a portion of the filesystem.**
- C - Failure of one namenode causes loss of some metadata availability from the entire filesystem.
- D - Each datanode registers with each namenode.

Q 23 - The main goal of HDFS High availability is

- A - Faster creation of the replicas of primary namenode.
- B - To reduce the cycle time required to bring back a new primary namenode after existing primary fails.**
- C - Prevent data loss due to failure of primary namenode.
- D - Prevent the primary namenode from becoming single point of failure.

Q 24 - As part of the HDFS high availability a pair of primary namenodes are configured. What is true for them?

- A - When a client request comes, one of them chosen at random serves the request.
- B - One of them is active while the other one remains powered off.
- C - Datanodes send block reports to only one of the namenodes.
- D - The standby node takes periodic checkpoints of active namenode's namespace.**

Q 25 - Zookeeper ensures that

- A - All the namenodes are actively serving the client requests
- B - Only one namenode is actively serving the client requests**
- C - A failover is triggered when any of the datanode fails.
- D - A failover can not be started by hadoop administrator.

Q 26 - Under Hadoop High Availability, Fencing means

- A - Preventing a previously active namenode from start running again.
- B - Preventing the start of a failover in the event of network failure with the active namenode.

C - Preventing the power down to the previously active namenode.

D - Preventing a previously active namenode from writing to the edit log.

Q 27 - Which of the following is not a fencing mechanism for a previously active namenode?

A - Disabling its network port via a remote management command.

B - Revoking its access to shared storage directory.

C - Formatting its disk drive.

D - STONITH

Q 28 - The property used to set the default filesystem for Hadoop in core-site.xml is-

A - filesystem.default

B - fs.default

C - fs.defaultFS

D - hdfs.default

Q 29 - The default replication factor for HDFS file system in hadoop is

A - 1

B - 2

C - 3

D - 4

Q 30 - When running on a pseudo distributed mode the replication factor is set to

A - 2

B - 1

C - 0

D - 3

Q 31 - For a HDFS directory the replication factor RF is

A - same as the RF of the files in that directory

B - Zero

C - 3

D - Does not apply.

Q 32 - The following is not permitted on HDFS files

A - Deleting

B - Renaming

C - Moving

D - Executing.

ANSWER SHEET

Question Number	Answer Key
------------------------	-------------------

1	C
---	---

2	A
---	---

3	D
---	---

4	B
---	---

5	B
---	---

6	C
---	---

7	C
---	---

8	B
---	---

9	C
---	---

10	D
----	---

11	B
----	---

12	B
----	---

13	A
----	---

14	B
----	---

15	A
----	---

16	C
----	---

17	A
----	---

18	D
----	---

19	A
----	---

20	C
----	---

21	C
----	---

22	B
----	---

23	B
----	---

24	D
----	---

25	B
----	---

26	D
----	---

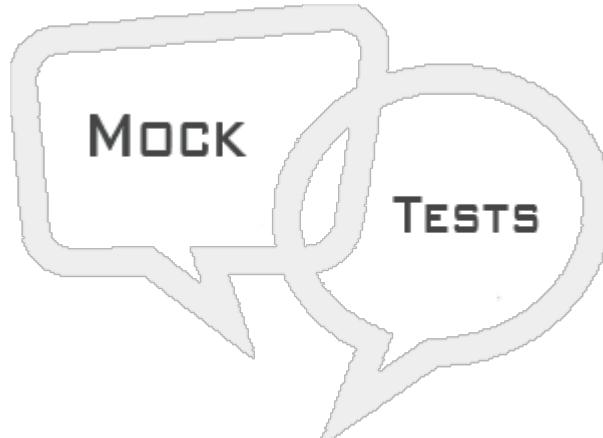
27	C
----	---

- | | |
|----|---|
| 28 | B |
| 29 | C |
| 30 | B |
| 31 | D |
| 32 | D |

Loading [MathJax]/jax/output/HTML-CSS/jax.js

HADOOP MOCK TEST

This section presents you various set of Mock Tests related to **Hadoop Framework**. You can download these sample mock tests at your local machine and solve offline at your convenience. Every mock test is supplied with a mock test key to let you verify the final score and grade yourself.



HADOOP MOCK TEST II

Q 1 - HDFS can be accessed over HTTP using

- A - viewfs URI scheme
- B - webhdfs URI scheme**
- C - wasb URI scheme
- D - HDFS ftp

Q 2 - What is true about HDFS?

- A - HDFS filesystem can be mounted on a local client's Filesystem using NFS.**
- B - HDFS filesystem can never be mounted on a local client's Filesystem.
- C - You can edit a existing record in HDFS file which is already mounted using NFS.
- D - You cannot append to a HDFS file which is mounted using NFS.

Q 3 - The client reading the data from HDFS filesystem in Hadoop

- A - gets the data from the namenode
- B - gets the block location from the datanode
- C - gets only the block locations form the namenode**
- D - gets both the data and block location from the namenode

Q 4 - Which scenario demands highest bandwidth for data transfer between nodes in Hadoop?

- A - Different nodes on the same rack

B - Nodes on different racks in the same data center.

C - Nodes in different data centers

D - Data on the same node.

Q 5 - The current block location of HDFS where data is being written to,

A - is visible to the client requesting for it.

B - Block locations are never visible to client requests.

C - May or may not be visible to the reader.

D - becomes visible only after the buffered data is committed.

Q 6 - Which of this is not a scheduler options available with YARN?

A - Optimal Scheduler

B - FIFO scheduler

C - Capacity scheduler

D - Fair scheduler

Q 7 - Which of the following is not a Hadoop operation mode?

A - Pseudo distributed mode

B - Globally distributed mode

C - Stand alone mode

D - Fully-Distributed mode

Q 8 - The difference between standalone and pseudo-distributed mode is

A - Stand alone cannot use map reduce

B - Stand alone has a single java process running in it.

C - Pseudo distributed mode does not use HDFS

D - Pseudo distributed mode needs two or more physical machines.

Q 9 - The hadoop frame work is written in

A - C++

B - Python

C - Java

D - GO

Q 10 - The hdfs command to create the copy of a file from a local system is

A - CopyFromLocal

B - copyfromlocal

C - CopyLocal

D - copyFromLocal

Q 11 - The hadfs command put is used to

A - Copy files from local file system to HDFS.

B - Copy files or directories from local file system to HDFS.

C - Copy files from from HDFS to local filesystem.

D - Copy files or directories from HDFS to local filesystem.

Q 12 - Underreplication in HDFS means-

A - No replication is happening in the data nodes.

B - Replication process is very slow in the data nodes.

C - The frequency of replication in data nodes is very low.

D - The number of replicated copies is less than as specified by the replication factor.

Q 13 - When the namenode finds that some blocks are over replicated, it

A - Stops the replication job in the entire hdfs file system.

B - It slows down the replication process for those blocks

C - It deletes the extra blocks.

D - It leaves the extra blocks as it is.

Q 14 - Which of the below property gets configured on core-site.xml ?

A - Replication factor

B - Directory names to store hdfs files.

C - Host and port where MapReduce task runs.

D - Java Environment variables.

Q 15 - Which of the below property gets configured on hdfs-site.xml ?

A - Replication factor

B - Directory names to store hdfs files.

C - Host and port where MapReduce task runs.

D - Java Environment variables.

Q 16 - Which of the below property gets configured on mapred-site.xml ?

- A - Replication factor
- B - Directory names to store hdfs files.
- C - Host and port where MapReduce task runs.**
- D - Java Environment variables.

Q 17 - Which of the below property gets configured on hadoop-env.sh?

- A - Replication factor
- B - Directory names to store hdfs files
- C - Host and port where MapReduce task runs
- D - Java Environment variables.**

Q 18 - The command to check if Hadoop is up and running is –

- A - Jsp
- B - Jps**
- C - Hadoop fs -test
- D - None

Q 19 - The information mapping data blocks with their corresponding files is stored in

- A - Data node
- B - Job Tracker
- C - Task Tracker
- D - Namenode**

Q 20 - The file in Namenode which stores the information mapping the data block location with file name is –

- A - dfsimage
- B - nameimage
- C - fsimage**
- D - image

Q 21 - The namenode knows that the datanode is active using a mechanism known as

- A - heartbeats**
- B - datapulse
- C - h-signal

D - Active-pulse

Q 22 - The nature of hardware for the namenode should be

- A - Superior than commodity grade
- B - Commodity grade
- C - Does not matter
- D - Just have more Ram than each of the data nodes

Q 23 - In Hadoop, Snappy and LZO are examples of

- A - Mechanisms of file transport between data nodes
- B - Mechanisms of data compression
- C - Mechanisms of data Replication
- D - Mechanisms of Data synchronization

Q 24 - Which of the below apache system deals with ingesting streaming data to hadoop

- A - Ozie
- B - Kafka
- C - Flume
- D - Hive

Q 25 - The input split used in MapReduce indicates

- A - The average size of the data blocks used as input for the program
- B - The location details of where the first whole record in a block begins and the last whole record in the block ends.
- C - Splitting the input data to a MapReduce program into a size already configured in the mapred-site.xml
- D - None of these

Q 26 - The output of a mapper task is

- A - The Key-value pair of all the records of the dataset.
- B - The Key-value pair of all the records from the input split processed by the mapper
- C - Only the sorted Keys from the input split
- D - The number of rows processed by the mapper task.

Q 27 - The role of a Journal node is to

- A - Report the location of the blocks in a data node
- B - Report the edit log information of the blocks in the data node.**
- C - Report the Schedules when the jobs are going to run
- D - Report the activity of various components handled by resource manager

Q 28 - The Zookeeper

- A - Detects the failure of the namenode and elects a new namenode.**
- B - Detects the failure of datanodes and elects a new datanode.
- C - Prevents the hardware from overheating by shutting them down.
- D - Maintains a list of all the components IP address of the Hadoop cluster.

Q 29 - If the IP address or hostname of a datanode changes

- A - The namenode updates the mapping between file name and block name
- B - The namenode need not update mapping between file name and block name**
- C - The data in that data node is lost forever
- D - The namenode has to be restarted

Q 30 - When a client contacts the namenode for accessing a file, the namenode responds with

- A - Size of the file requested.
- B - Block ID of the file requested.
- C - Block ID and hostname of any one of the data nodes containing that block.
- D - Block ID and hostname of all the data nodes containing that block.**

Q 31 - HDFS stands for

- A - Highly distributed file system.
- B - Hadoop directed file system
- C - Highly distributed file shell
- D - Hadoop distributed file system.**

Q 32 - The Hadoop tool used for uniformly spreading the data across the data nodes is named –

- A - Scheduler
- B - Balancer**
- C - Spreader

Q 33 - In the secondary namenode the amount of memory needed is

- A - Similar to that of primary node
- B - Should be at least half of the primary node
- C - Must be double of that of primary node
- D - Depends only on the number of data nodes it is going to handle

ANSWER SHEET

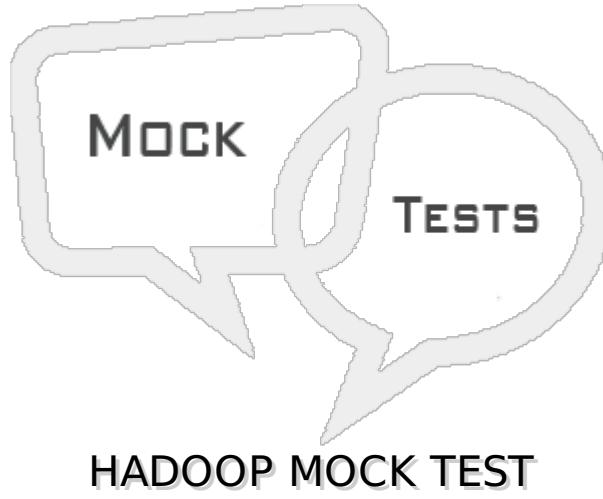
Question Number Answer Key

1	B
2	A
3	C
4	C
5	D
6	A
7	B
8	B
9	C
10	D
11	B
12	D
13	C
14	B
15	A
16	C
17	D
18	B
19	D
20	C
21	A
22	A
23	B
24	C

25	B
26	B
27	B
28	A
29	B
30	D
31	D
32	B
33	A

HADOOP MOCK TEST

This section presents you various set of Mock Tests related to **Hadoop Framework**. You can download these sample mock tests at your local machine and solve offline at your convenience. Every mock test is supplied with a mock test key to let you verify the final score and grade yourself.



HADOOP MOCK TEST

Q 1 - The purpose of checkpoint node in a Hadoop cluster is to

- A - Check if the namenode is active
- B - Check if the fsimage file is in sync between namenode and secondary namenode
- C - Merges the fsimage and edit log and uploads it back to active namenode.**
- D - Check which data nodes are unreachable.

Q 2 - When a backup node is used in a cluster there is no need of

- A - Check point node**
- B - Secondary name node
- C - Secondary data node
- D - Rack awareness

Q 3 - Rack awareness in name node means

- A - It is aware how many racks are available in the cluster
- B - It is aware of the mapping between the node and the rack**
- C - It is aware of the number of nodes in each of the rack
- D - It is aware which data nodes are unavailable in the cluster.

Q 4 - When a machine is declared as a datanode, the disk space in it

- A - Can be used only for HDFS storage
- B - Can be used for both HDFS and non-HDFS storage**

C - Cannot be accessed by non-hadoop commands

D - cannot store text files.

Q 5 - When a file in HDFS is deleted by a user

A - it is lost forever

B - It goes to trash if configured.

C - It becomes hidden from the user but stays in the file system

D - File sin HDFS cannot be deleted

Q 6 - The source of HDFS architecture in Hadoop originated as

A - Google distributed filesystem

B - Yahoo distributed filesystem

C - Facebook distributed filesystem

D - Azure distributed filesystem

Q 7 - The inter process communication between different nodes in Hadoop uses

A - REST API

B - RPC

C - RMI

D - IP Exchange

Q 8 - The type of data Hadoop can deal with is

A - Structred

B - Semi-structured

C - Unstructured

D - All of the above

Q 9 - YARN stands for

A - Yahoo's another resource name

B - Yet another resource negotiator

C - Yahoo's archived Resource names

D - Yet another resource need.

Q 10 - The fully distributed mode of installation without virtualization needs a minimum of

A - 2 physical machines

B - 3 Physical machines

C - 4 Physical machines

D - 1 Physical machine

Q 11 - Running Start-dfs.sh results in

A - Starting namenode and datanode

B - Starting namenode only

C - Starting datanode only

D - Starting namenode and resource manager

Q 12 - Which of the following is not a goal of HDFS?

A - Fault detection and recovery

B - Handle huge dataset

C - Prevent deletion of data

D - Provide high network bandwidth for data movement

Q 13 - The command “hadoop fs -test -z URI “ gives the result 0 if

A - if the path is a directory

B - if the path is a file

C - if the path is not empty

D - if the file is zero length

Q 14 - In HDFS the files cannot be

A - read

B - deleted

C - executed

D - Archived

Q 15 - hadoop fs -expunge

A - Gives the list of datanodes

B - Used to delete a file

C - Used to exchange a file between two datanodes.

D - Empties the trash.

Q 16 - All the files in a directory in HDFS can be merged together using

- A - getmerge
- B - putmerge
- C - remerge
- D - mergeall

Q 17 - The replication factor of a file in HDFS can be changed using

- A - changerep
- B - rerep
- C - setrep
- D - xrep

Q 18 - The command used to copy a directory from one node to another in HDFS is

- A - rcp
- B - dcp
- C - drcp
- D - distcp

Q 19 - The archive file created in Hadoop always has the extension of

- A - .hrc
- B - .har
- C - .hrh
- D - .hrar

Q 20 - To unarchive an already archived file in hadoop use the command

- A - unrar
- B - unhar
- C - cp
- D - cphar

Q 21 - The data from a remote hadoop cluster can

- A - not be read by another hadoop cluster
- B - be read using http
- C - be read using hhttp

D - be read suing hftp

Q 22 - The purpose of starting namenode in the recovery mode is to

- A - Recover a failed namenode
- B - Recover a failed datanode
- C - Recover data from one of the metadata storage locations
- D - Recover data when there is only one metadata storage location**

Q 23 - When you increase the number of files stored in HDFS, The memory required by namenode

- A - Increases**
- B - Decreases
- C - Remains unchanged
- D - May increase or decrease

Q 24 - If we increase the size of files stored in HDFS without increasing the number of files, then the memory required by namenode

- A - Decreases**
- B - Increases
- C - Remains unchanged
- D - May or may not increase

Q 25 - The current limiting factor to the size of a hadoop cluster is

- A - Excess heat generated in data center
- B - Upper limit of the network bandwidth
- C - Upper limit of the RAM in namenode**
- D - 4000 data nodes

Q 26 - The decommission feature in hadoop is used for

- A - Decommissioning the namenode
- B - Decommissioning the data nodes**
- C - Decommissioning the secondary namenode.
- D - Decommissioning the entire Hadoop cluster.

Q 27 - You can reserve the amount of disk usage in a data node by configuring the dfs.datanode.du.reserved in which of the following file

A - Hdfs-site.xml

B - Hdfs-defaukt.xml

C - Core-site.xml

D - Mapred-site.xml

Q 28 - The namenode loses its only copy of fsimage file. We can recover this from

A - Datanodes

B - Secondary namenode

C - Checkpoint node

D - Never

Q 29 - In a HDFS system with block size 64MB we store a file which is less than 64MB. Which of the following is true?

A - The file will consume 64MB

B - The file will consume more than 64MB

C - The file will consume less than 64MB.

D - Can not be predicted.

Q 30 - A running job in hadoop can

A - Be killed with a command

B - Can never be killed with a command

C - Can be killed only by shutting down the name node

D - Be paused and run again

Q 31 - The number of tasks a task tracker can accept depends on

A - Maximum memory available in the node

B - Not limited

C - Number of slots configured in it

D - As decided by the jobTracker

Q 32 - When a jobTracker schedules a task is first looks for

A - A node with empty slot in the same rack as datanode

B - Any node on the same rack as the datanode

C - Any node on the rack adjacent to rack of the datanode

D - Just any node in the cluster

Q 33 - The heartbeat signal are sent from

A - JObtracker to Tasktracker

B - Tasktracker to Job tracker

C - Jobtracker to namenode

D - Tasktracker to namenode

ANSWER SHEET

Question Number Answer Key

1 C

2 A

3 B

4 B

5 B

6 A

7 B

8 D

9 B

10 A

11 A

12 C

13 D

14 C

15 D

16 A

17 C

18 D

19 B

20 C

21 D

22 D

23 A

24 A

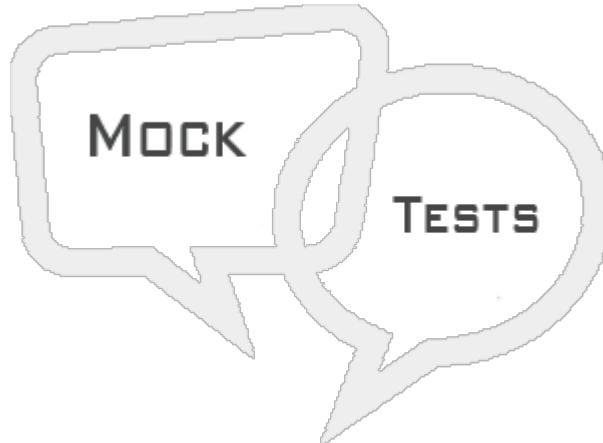
25 C

- | | |
|----|---|
| 26 | B |
| 27 | A |
| 28 | C |
| 29 | C |
| 30 | A |
| 31 | C |
| 32 | A |
| 33 | B |

Loading [MathJax]/jax/output/HTML-CSS/jax.js

HADOOP MOCK TEST

This section presents you various set of Mock Tests related to **Hadoop Framework**. You can download these sample mock tests at your local machine and solve offline at your convenience. Every mock test is supplied with a mock test key to let you verify the final score and grade yourself.



HADOOP MOCK TEST IV

Q 1 - When a jobTracker schedules a task is first looks for

- A - A node with empty slot in the same rack as datanode
- B - Any node on the same rack as the datanode
- C - Any node on the rack adjacent to rack of the datanode
- D - Just any node in the cluster

Q 2 - The heartbeat signal are sent from

- A - JOBtracker to Tasktracker
- B - Tasktracker to Job tracker**
- C - Jobtracker to namenode
- D - Tasktracker to namenode

Q 3 - Job tracker runs on

- A - Namenode**
- B - Datanode
- C - Secondary namenode
- D - Secondary datanode

Q 4 - Which of the following is not a scheduling option available in YARN

- A - Balanced scheduler**
- B - Fair scheduler

C - Capacity scheduler

D - FiFO schesduler.

Q 5 - What is the default input format?

A - The default input format is xml. Developer can specify other input formats as appropriate if xml is not the correct input.

B - There is no default input format. The input format always should be specified.

C - The default input format is a sequence file format. The data needs to be preprocessed before using the default input format.

D - The default input format is TextInputFormat with byte offset as a key and entire line as a value.

Q 6 - Which one is not one of the big data feature?

A - Velocity

B - Veracity

C - volume

D - variety

Q 7 - Which technology is used to store data in Hadoop?

A - HBase

B - Avro

C - Sqoop

D - Zookeeper

Q 8 - Which technology is used to serialize the data in Hadoop?

A - HBase

B - Avro

C - Sqoop

D - Zookeeper

Q 9 - Which technology is used to import and export data in Hadoop?

A - HBase

B - Avro

C - Sqoop

D - Zookeeper

Q 10 - Which of the following technologies is a document store database?

- A - HBase
- B - Hive
- C - Cassandra
- D - CouchDB**

Q 11 - Which one of the following is not true regarding to Hadoop?

- A - It is a distributed framework.
- B - The main algorithm used in it is Map Reduce
- C - It runs with commodity hard ware
- D - All are true**

Q 12 - Which one of the following stores data?

- A - Name node
- B - Data node**
- C - Master node
- D - None of these

Q 13 - Which one of the following nodes manages other nodes?

- A - Name node**
- B - Data node
- C - slave node
- D - None of these

Q 14 - What is AVRO?

- A - Avro is a java serialization library.**
- B - Avro is a java compression library.
- C - Avro is a java library that create split table files.
- D - None of these answers are correct.

Q 15 - Can you run Map - Reduce jobs directly on Avro data?

- A - Yes, Avro was specifically designed for data processing via Map-Reduce.**
- B - Yes, but additional extensive coding is required.
- C - No, Avro was specifically designed for data storage only.
- D - Avro specifies metadata that allows easier data access. This data cannot be used as part of

map-reduce execution, rather input specification only.

Q 16 - What is distributed cache?

- A - The distributed cache is special component on name node that will cache frequently used data for faster client response. It is used during reduce step.
- B - The distributed cache is special component on data node that will cache frequently used data for faster client response. It is used during map step.
- C - The distributed cache is a component that caches java objects.
- D - The distributed cache is a component that allows developers to deploy jars for Map-Reduce processing.

Q 17 - What is writable?

- A - Writable is a java interface that needs to be implemented for streaming data to remote servers.
- B - Writable is a java interface that needs to be implemented for HDFS writes.
- C - Writable is a java interface that needs to be implemented for MapReduce processing.
- D - None of these answers are correct.

Q 18 - What is HBASE?

- A - Hbase is separate set of the Java API for Hadoop cluster.
- B - Hbase is a part of the Apache Hadoop project that provides interface for scanning large amount of data using Hadoop infrastructure.
- D - HBase is a part of the Apache Hadoop project that provides a SQL like interface for data processing.

Q 19 - How does Hadoop process large volumes of data?

- A - Hadoop uses a lot of machines in parallel. This optimizes data processing.
- B - Hadoop was specifically designed to process large amount of data by taking advantage of MPP hardware.
- C - Hadoop ships the code to the data instead of sending the data to the code.
- D - Hadoop uses sophisticated caching techniques on name node to speed processing of data.

Q 20 - When using HDFS, what occurs when a file is deleted from the command line?

- A - It is permanently deleted if trash is enabled.
- B - It is placed into a trash directory common to all users for that cluster.
- C - It is permanently deleted and the file attributes are recorded in a log file.
- D - It is moved into the trash directory of the user who deleted it if trash is enabled.

Q 21 - When archiving Hadoop files, which of the following statements are true?

Choose two answers

1. Archived files will display with the extension .arc.
2. Many small files will become fewer large files.
3. MapReduce processes the original files names even after files are archived.
4. Archived files must be UN archived for HDFS and MapReduce to access the original, small files.
5. Archive is intended for files that need to be saved but no longer accessed by HDFS.

A - 1 & 3

B - 2 & 3

C - 2 & 4

D - 3 & 4

Q 22 - When writing data to HDFS what is true if the replication factor is three?

Choose 2 answers

1. Data is written to DataNodes on three separate racks *if RackAware*.
2. The Data is stored on each DataNode with a separate file which contains a checksum value.
3. Data is written to blocks on three different DataNodes.
4. The Client is returned with a success upon the successful writing of the first block and checksum check.

A - 1 & 3

B - 2 & 3

C - 3 & 4

D - 1 & 4

Q 23 - Which of the following are among the duties of the Data Nodes in HDFS?

A - Maintain the file system tree and metadata for all files and directories.

B - None of the options is correct.

C - Control the execution of an individual map task or a reduce task.

D - Store and retrieve blocks when told to by clients or the NameNode.

E - Manage the file system namespace.

Q 24 - Which of the following components retrieves the input splits directly from HDFS to determine the number of map tasks?

A - The NameNode.

B - The TaskTrackers.

C - The JobClient.

D - The JobTracker.

E - None of the options is correct.

Q 25 - The org.apache.hadoop.io.Writable interface declares which two methods?

Choose2answers.

1. **public void readFieldsDataInput.**
2. **public void readDataInput.**
3. **public void writeFieldsDataOutput.**
4. **public void writeDataOutput.**

A - 1 & 4

B - 2 & 3

C - 3 & 4

D - 2 & 4

Q 26 - Which one of the following statements is true regarding <key,value> pairs of a MapReduce job?

A - A key class must implement Writable.

B - A key class must implement WritableComparable.

C - A value class must implement WritableComparable.

D - A value class must extend WritableComparable.

Q 27 - Which one of the following statements is false regarding the Distributed Cache?

A - The Hadoop framework will ensure that any files in the Distributed Cache are distributed to all map and reduce tasks.

B - The files in the cache can be text files, or they can be archive files like zip and JAR files.

C - Disk I/O is avoided because data in the cache is stored in memory.

D - The Hadoop framework will copy the files in the Distributed Cache on to the slave node before any tasks for the job are executed on that node.

Q 28 - Which one of the following is not a main component of HBase?

A - Region Server.

B - Nagios.

C - ZooKeeper.

D - Master Server.

Q 29 - Which of the following is false about RawComparator ?

A - Compare the keys by byte.

B - Performance can be improved in sort and shuffle phase by using RawComparator.

C - Intermediary keys are deserialized to perform a comparison.

Q 30 - Which demon is responsible for replication of data in Hadoop?

A - HDFS.

B - Task Tracker.

C - Job Tracker.

D - Name Node.

E - Data Node.

Q 31 - Keys from the output of shuffle and sort implement which of the following interface?

A - Writable.

B - WritableComparable.

C - Configurable.

D - ComparableWritable.

E - Comparable.

Q 32 - In order to apply a combiner, what is one property that has to be satisfied by the values emitted from the mapper?

A - Combiner can be applied always to any data

B - Output of the mapper and output of the combiner has to be same key value pair and they can be heterogeneous

C - Output of the mapper and output of the combiner has to be same key value pair. Only if the values satisfy associative and commutative property it can be done.

ANSWER SHEET

Question Number Answer Key

1 A

2 B

3 A

4 A

5 D

6 B

7 A

8	B
9	C
10	D
11	D
12	B
13	A
14	A
15	A
16	B
17	C
18	B
19	C
20	C
21	B
22	C
23	D
24	D
25	A
26	B
27	C
28	B
29	C
30	D
31	B
32	C

Loading [MathJax]/jax/output/HTML-CSS/jax.js

- (A) Structured
- (B) Unstructured
- (C) Processed
- (D) Semi-Structured

Answer

C

MCQ No - 19

Big Data applications benefit the media and entertainment industry by

- (A) Predicting what the audience wants
- (B) Ad targeting
- (C) Scheduling optimization
- (D) All of the above

Answer

D

MCQ No - 20

The feature of big data that refers to the quality of the stored data is

-
- (A) Variety
 - (B) Volume
 - (C) Variability
 - (D) Veracity

Answer

D

Question 1

What is the difference between interval/ratio and ordinal variables?

- a)** The distance between categories is equal across the range of interval/ratio data.

Feedback:

The data that we gather varies from person to person. People are of different ages, have different income levels and prefer to do some things more than other people. We call these things variables just because their values vary from person to person. Analysis of quantitative data starts by trying to understand what kinds of variables we are dealing with. A person's age is an example of an interval/ratio variable, because ages are measured in years. We can do a lot of statistical analysis on this kind of variable because the interval (one year) is the same for everybody in our data-set. Some variables are called 'dichotomous', meaning all possible answers are of one of two types (male/female, for example). We call those variables 'nominal', which we can, literally, only "name", like many types of job occupation, for example. Finally, we refer to some variables as 'ordinal', which means we can only place the values in an order of first, second, third and so on, without considering the gap between the first and second, or whether it was the same as between second and third. Apart from dichotomous variables, all others can be rank-ordered.

Question 2

What is the difference between a bar chart and a histogram?

- c)** There are no gaps between the bars on a histogram.

Feedback:

Histograms are used to display interval/ratio variables, which involve a continuous range of values, and so there are no gaps between the bars that represent each category. Bar charts, on the other hand, display nominal or ordinal data, which fall into discrete categories.

Question 3

What does the term 'outlier' mean?

- d)** An extreme value at either end of a distribution

Feedback:

When we calculate a simple average, the 'arithmetic mean', we have to remember that a wide range of values can give the same average as a narrow range and that extreme values could make a simple average fairly meaningless. These values are called 'outliers', extremely high or low values in a distribution that threaten to skew the results. The 'median' is useful, in this regard, because it simply identifies the mid-point in a whole array of values, giving us a measure of the significance of the arithmetic mean.

Question 4

What is the function of a contingency table, in the context of bivariate analysis?

Correct answer:

- b) It summarizes the frequencies of two variables so that they can be compared.

Feedback:

'Bivariate' analysis means that we are analysing two variables together, usually to see if any co-relation exists between them. There are various techniques available for this, one of which is a contingency table. This technique is principally used to compare nominal variables with another type, where the frequencies (in numbers or percentages) of the two different variables are simultaneously analysed to identify patterns of association between them.

Question 5

If there were a perfect positive correlation between two interval/ratio variables, the Pearson's *r* test would give a correlation coefficient of:

Correct answer:

- b) +1

Feedback:

A coefficient is a measure of the degree to which two sets of numbers co-relate. If the variables always move in 'lock-step' with each other, we call that a 'perfect' correlation. Sometimes the variables move in the same direction as each other, a 'positive' correlation and sometimes in the opposite direction, a 'negative' correlation. Pearson's *r* test gives an answer of +1 when there is a perfect positive correlation between interval/ratio variables and -1 when there is a perfect negative correlation between them.

Question 6

What is the name of the test that is used to assess the relationship between two ordinal variables?

Correct answer:

- a) Spearman's *rho*

Feedback:

Pearson's *r* test is extremely valuable but limited to assessing correlations between interval/ratio variables. Spearman's *rho* test is a very similar technique which can be used on pairs of variables when either both are ordinal or one is ordinal and the other is interval/ratio. The result will lie between -1 and +1, indicating the range of possible correlation, from perfectly negative to perfectly positive. The phi coefficient is used for

dichotomous variables and Cramer's V is a test of the strength of the relationship between nominal variables. Chi square, in brief, tests for the likelihood of relationships existing through mere chance, so is usually used in conjunction with the tests discussed in this question.

Question 7

When might it be appropriate to conduct a multivariate analysis test?

Correct answer:

- d)** All of the above.

Feedback:

Multivariate analysis involves the analysis of three or more variables, and tends to be used when we have reason to suspect the nature of the relationship between two variables. Two variables can, indeed, be related to each other but perhaps in a more complex way than appears at first sight. Perhaps when a number of factors co-exist the relationship between any two of them is strong. Multivariate analysis enables us to test for many types of cross-relationships between a number of variables, at once.

Question 8

What is meant by a "spurious" relationship between two variables?

Correct answer:

- c)** A relationship that appears to be true because each variable is related to a third one.

Feedback:

One of the conditions under which it is appropriate to use multivariate analysis is when the relationship between two variables might be spurious: this means that the relationship, which seemed to exist, doesn't exist in reality. A third variable turns out, perhaps, to be responsible for the variation in both sets of values, and so they are not really related to each other, so their relationship was "spurious".

Question 9

A test of statistical significance indicates how confident the researcher is about:

Correct answer:

- d)** generalising their findings from the sample to the population.

Feedback:

Tests of statistical significance allow the researcher to estimate how confident they can be that there is a real relationship between the variables they are studying and thus that their

results can be generalised from the sample to the target population.

Question 10

Setting the p level at 0.01 increases the chances of making a:

Correct answer:

b) Type II error

Feedback:

The p value represents the level of probability that an apparently significant relationship between variables was really just due to chance. If p is set at 0.01, this means that we would expect such a result in only 1 in 100 cases. This is a very stringent level, and while it means that the researcher can be more confident about a significant result if they find one, it also increases the chance of making a Type II error: confirming the null hypothesis when it should be rejected.

D. Y. Patil College of Engineering
Akurdi, Pune-44.
Department of Computer Engineering

Class : BE Div: A+B Subject : Data Analytics
 Academic Year : 2020-21 Sem : I

UNIT 1

Q No	Question Statement	Options	Correct Answer
1	Select the data types used for Big Data .	1.Structured 2. Semistructured 3. Quasistructured 4.Unstructured 5. All of the above	5
2	CCTV coverage is example of _____ data	1.Structured 2. Semistructured 3. Quasistructured 4.Unstructured 5. All of the above	2
3	Players in the big data ecosystem are:	1.Data Devices 2. Data collectors 3. Data aggregators 4.Data users and buyers 5.All of the above	5
4	Data users are benefited from Data Collector and Data Aggregator.	1.TRUE 2.FALSE	2
5	Match the following a)Structured Data 1)XML data b)Semistructured data 2)RDBMS c)Quasistructured data 3)Images, video d)Unstructured data 4)Clickstream	a)Structured Data 1)XML data b)Semistructured data 2)RDBMS c)Quasistructured data 3)Images, video d)Unstructured data 4)Clickstream	a-2 b-1 c-4 d-3
6	Select Key roles for New big data ecosystem are 1.Deep analytical talent 2. Data Savvy Professionals 3. _____	1.DatavScientist 2.Statician 3.Mathematician 4.Technology and data enablers	4
7	Apply your knowledge to identify skills required for Technology and Data enablers a)Managing large data architecture	1.a 2.b 3.a and b	3

	b) Administrating analytical sandboxes	4.cant say	
8	CRISP-DM stand for	1.Cross Industry Standard Process for Data Mining 2.Critical Industry Standard Process for Data Mining 3. Complex Industry Standard Process for Data Mining 4.None of the mentioned	1
9	Phase1 in data analytics lifecycle is	1.Data Discovery 2.Data Preparation 3.Model Planning 4.Communication	1
10	Understand the Phases in data analytics lifecycle and state that phases in data analytics lifecycle are working on	1.Forward Approach 2.Backward Approach 3.Iterative Approach 4.None of the mentioned	3
11	Data scientist is the one who_____	1Creates dashboards, reports 2. Provide and configure database environment 3. Design ,implement and deploy model 4. Provides domain expertise based on KPI	3
12	In Model Planning phase[Phase3] , data exploration is done to	1. Understand the relationships among the variables to inform selection of the variables 2. Preprocess and clean data to generate quality data 3. Discover the data from various data sources	1
13	Which of the phases in data analytics life cycle consume more time of data science team?	1.Data Discovery and Data Preparation 2.Model Planning and Model Building 3. Model Planning and Operationalize 4. Can't say	1
14	Which among the following is not commercial tool for model building?	1. IBM -SPSS Modeler 2. SAS Enterprise miner 3. AlpineMiner	4

		4. WEKA	
15	Case Study :GINA describes	<ol style="list-style-type: none"> 1. Current analytical architecture 2. Big data ecosystem 3. Data analytics lifecycle 4. Sources of big data deluge 	3

Unit 2

Q No	Question Statement	Options	Correct Answer
1	_____ a measure of variation of scores about the mean	<ol style="list-style-type: none"> 1. Standard Deviation 2. Variance 3. Can't Say 	1
2	Parametric hypothesis test procedures involve	<ol style="list-style-type: none"> 1. Stringent assumptions 2. Population parameters 3. Both 1 and 2 4. none of the mentioned 	3
3	Nonparametric hypothesis test procedure examples are	<ol style="list-style-type: none"> 1. Z test 2. t-test 3. Wilcoxon Rank Sum Test 4. Both Z test and t-Test 5. Z test and t-Test and Wilcoxon Rank Sum Test 	4
4	The test statistic for the Wilcoxon Rank Sum Test is based on which of the following?	<ol style="list-style-type: none"> 1. The sum of the ranks of the pooled data 2. The sum of the ranks of the smaller sample in the pooled data 3. The correlation of the ranks of the variables 4. None of the above 	2
5	Situation when we reject null hypothesis when it is true is _____	<ol style="list-style-type: none"> 1. Type -I Error 2. Type -II Error 3. Type-III Error 4. None of the mentioned 	1

6	Consider the dataset 1 1.50 2 1 0.50 calculate variance	1.0.325 2.0.425 3.0.125 4.0.225	1
7	The Wilcoxon rank sum test assumes that the population distribution in the difference of paired samples	1. follows the normal distribution. 2. follows the Student's t distribution. 3. follows the chi-square distribution. 4. is symmetrical.	1
8	Statistical methods are used throughout the data analytics lifecycle.	1.TRUE 2.FALSE	1
9	In linear regression $y=mx+c$, identify dependent variable.	1.y 2.x 3.m 4.c	1
10	In statistics μ is called as	1. Mean of population 2. Mean of samples 3. standard deviation for population 4. standard deviation for samples	1
11	_____ is measure of how data points differ from the mean ?	1.Variance 2.Mode 3. Standard deviation 4.can't say	1
12	For factory 1, the rates are <ul style="list-style-type: none">• 71, 82, 77, 92, 88.• For factory 2, the rates are<ul style="list-style-type: none">• 85, 82, 94 & 97. For Wilcoxon Rank sum test calculate the rank sums for Factory1 and Factory 2 ?	1.Factory 1 : 19.5 Factory2:25.5 2.Factory 1 : 25.5 Factory2:19.5 3.Factory 1 : 19 Factory2:25 4. Factory 1 : 25 Factory2:19	1

13	<p>Given {2,4,10,12,3,20,30,11,25}, k=2</p> <p>Randomly assign means: $m_1=3, m_2=4$</p> <p>Apply K means clustering algorithm and find the elements in cluster K1 and K2</p>	<p>1. $K_1=\{2,3,4,10,11,12\}$, $K_2=\{20,30,25\}$,</p> <p>2. $K_1=\{2,3\}$, $K_2=\{4,10,12,20,30,11,25\}$,</p> <p>3. $K_1=\{2,3,4\}$, $K_2=\{10,12,20,30,11,25\}$</p> <p>4. $K_1=\{2,3,4,10\}$, $K_2=\{12,20,3,11,25\}$,</p>	1
14	K Means algorithm uses	<p>1. Supervised learning 2. Unsupervised Learning 3. Semisupervised learning 4. None of the mentioned</p>	2
15	Variants of K means algorithm uses distance measures as	<p>1. Euclidian Distance 2. Manhattan Distance 3. Mahalonobias distance 4. Can't Say</p>	3

UNIT3

Q No	Question Statement	Options	Correct Answer
1	Apply Apriory property and confirm that if {A,B,C,D} is a frequent itemset, _____ candidate rules which is not possible	<p>1. $C \rightarrow A$ 2. $D \rightarrow ABCD$ 3. $A \rightarrow BC$ 4. $B \rightarrow ADC$</p>	4
2	Association Rule mining concludes that prefer	<p>1. High support and low confidence 2. Low support and high confidence</p>	2

		3. Low support and low confidence 4. High support and medium confidence	
3	Data set {brown, black, blue, green , red} is example of	1. Continuous attribute 2. Ordinal attribute 3. Numeric attribute 4. Nominal attribute	4
4	Interpret Association rule mining is A] Descriptive Method B]Unsupervised Learning method	1.A 2.B 3.A&B 4.Can't say	3
5	Use cases for Apriory algorithm include A]Items purchased in one transaction B]Set of hyperlinks clicked by a user in one session C]Prediction of income if education and experience is known.	5. A 6. B 7. A and B 8. C	3
6	Apriory property states that If an itemset is frequent, then any subset is _____	1.frequent 2.not frequent 3.sometimes frequent 4.can't say	1
7	_____ calculates influence that input (independent) variables have on the outcome (dependent) variable	1.Apriory Algorithm 2.Linear regression 3.clustering 4.Can't say	2
8	Simple regression assumes a _____ relationship between the input attribute and output attribute.	1 reciprocal 2.quadratic 3 inverse 4. linear	4
9	In linear regression modeling, the outcome variable is _____	1.Continuous 2.Categorial 3.Nominal 4.Can't say	1
10	_____ is defined as the difference between the observed outcome variables and the fitted value based on the OLS parameter estimates	1.Residuals 2.p value 3.cofidance interval; 4.Can't say	1

11	In logistic regression modeling, the outcome variable describes _____	1.Continuous 2.Categorical 3.Numerical 4.Can't say	2
12	Use Cases of Logistic regression include A. Probability a wireless customer switches carriers (churns) B. Restaurant predicts quantity of food needed ,if possible inputs – weather, day of week, etc. are known	1.A 2.B 3.A and B 4.Can't Say	1
13	_____ splits the dataset into training and test sets	1.Overfitting 2.Underfitting 3.Model fitting 4.N-fold cross validation	4
14	In linear regression $Y=b_0+b_1X_1+b_2X_2$, Identify dependent variable .	5. Y 6. X 7. B 8. Can't say	1
15	_____assume a linear additive function of the inputs variables .	5. Linear Regression model 6. Logistic regression model 7. Both Linear Regression and Logistic Regression 8. Can't Say	3

Question	Option 1	Option 2	Option 3	Option 4	ANSWER
1 Recognize meaning “ High entropy means that the partitions in classification are...”	Pure	Not pure	useful	useless	2
2 Which of the following statements about Naive Bayes is incorrect?	Attributes are equally important.	Attributes are statistically dependent of one another given the class value	Attributes are statistically independent of one another given the class value.	Attributes can be nominal or numeric	2
3 What are the axes of an ROC curve?	Vertical axis: % of true negatives; Horizontal axis: % of false negatives	Vertical axis: % of true positives; Horizontal axis: % of false positives	Vertical axis: % of false negatives; Horizontal axis: % of false positives	Vertical axis: % of false positives; Horizontal axis: % of true negatives	2
4 Discover Which of the following is NOT supervised learning?	PCA	Decision Tree	Linear Regression	Naive Bayesian	1
5 Define 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	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	None of the mentioned	3
6 Bayes rule can be used for:-	Solving queries	Increasing complexity	Answering probabilistic query	Decreasing complexity	3
7 Discover the disadvantage of decision trees?	Factor analysis	Decision trees are robust to outliers	Decision trees are prone to be overfit	None of the above	3
8 _____ refers to a model that can neither model the training data nor generalize to new data	good fitting	Overfitting	underfitting	all of the above	3
9 Naïve Bayes algorithm is based on _____ and used for solving classification problems.	Bayes Theorem	Candidate elimination algorithm	EM algorithm	None of the above	1
10 Naïve Bayes Algorithm is a ___ learning algorithm.	Supervised	Reinforcement	Unsupervised	None of these	1
11 Data used to optimize the parameter settings of a supervised learner model.	a. training	Test	Verification	Validation	4
12 What are tree based classifiers?	Classifiers which form a tree with each attribute at one level	Classifiers which perform series of condition checking with one attribute at a time	Both options except none	None of the options	3
13 Applications of Naïve Bayes Algorithm is/are	Spam filtration	Sentimental analysis	Classifying articles	All of the above	4
14 Select The benefit of Naïve Bayes:-	Naïve Bayes is one of the fast and easy ML algorithms to predict a class of datasets.	It is the most popular choice for text classification problems.	It can be used for Binary as well as Multi-class Classifications	All of the above	4
15 Basic algorithm for decision tree induction is designed using	Greedy Approach	Top Down Approach	Procedural Approach	Step by Step Approach	1

UNITS

	Question*	Option 1	Option 2	Option 3	Option 4	ANSWER
1	Which of the following is visualized using boxplot?	Collection Of Values	distribution of values	Frequency of Values	None	2
2	Which of the following is used for visualizing Comparison of relative point values?	Piechart	Bar chart	LineChart	None	2
3	Which of the following summarises a set of categorical/nominal data?	Piechart	Bar chart	LineChart	None	1
4	Which of the following is not a Free Visualization Tool?	R Shiny	Tableau Public	D3.js	QlikView	4
5	Which of the following functions in R, is used to represent 1-d plot of the data to an existing plot?	Plot(data)	barplot(data)	rug(data)	stem(data)	3
6	Which of the following refers to elements of data visualization that provide additional information but do not contribute to the data portion of the graphic?	Whiskers	Chart Junk	Scatters	Histogram	2
7	Which of the following chart junks is represented by large square blocks, but doesnot represent any specific meaning?	Chunky data points	Crowded axis labels	Horizontal grid lines	No context	1
8	Which of the following is not a benefit of Data Visualization?	It allows users to see several perspectives of data	Intepret vast amount of data	with all the data available, it is easy to find the nuances that can make a difference	Exploring trends within a database by letting analysts navigate through data and visually orient themselves to the patterns in the data	3
9	Select a scenario where Scatterplots are not suitable to be used.	to show the relationship between two variables	when looking for outliers or for understanding the distribution of your data.	You want clear and precise data points	You want a compact data visualization	3
10	Which of the following is a JavaScript library for manipulating data and creating web based visualization with standards, such as HTML, SVG and CSS?	D3.js	TIBCO	Rggobi	Tableau	1
11	Which of the following can be exposed and recognized easier with data visualization software?	Patterns	trends	correlations	ALL	4
12	Which of the following is expected to be provided by visualization tool with respect to interactive viisualization?	high latency	low latency	latency	None	2

UNIT6

	Question*	Option 1	Option 2	Option 3	Option 4	Correct Option
1	Which of the following tasks is not performed by Mahout package?	Collaborative Filtering	Clustering	Structuring and storing large datasets	Classification	3
2	Which of the following recommendation system is used in Mahout ?	Collaborative Filtering	Clustering	Content based filtering	Hybrid filtering	1
3	Geospatial applications such as delivery and traffic systems used to optimize the time to reach one or more destinations, are examples of which of the following ?	Key/Value stores	Document Stores	Graph Databases	Column family stores	3
4	In the final phase of the Data Analytics lifecycle, which of the following serve as the key output from the perspective of a Data Scientist?	Presentation for Analysts	Code and presentation for analyst	Technical Specs	Presentation for project sponsors	2
5	In the final phase of the Data Analytics lifecycle, which of the following serve as the key output from the perspective of a Database Administrator?	Presentation for Analysts	Code and presentation for analyst	Code and Technical Specs	Presentation for project sponsors	3
6	Which of the following deliverables are common in the presentation for Project sponsors and Analysts?	Project Goals	Main Findings	Model Description	All the options	4
7	Which of the following deliverables are not common in the presentation for Project sponsors and Analysts?	Model Details	Project Goals	Main Findings	Model Description	1
8	Supplementing recommendations with implications for the modeling or for deploying in a production environment. This is a requirement of which of the following stakeholder of a Data Analytics project?	Project Manager	Data Engineer	Database Administrator	Business Intelligence Analyst	4
9	Which of the following deliverables are not required in the presentation for Project sponsors?	Model Details	Project Goals	Main Findings	Model Description	1
10	Which of the following programming languages is used to write MAHOUT	C	C++	Scala	Python	3
11	Which of the following is not a data collection tool in the Hadoop Ecosystem?	Flume	Sqoop	Avro	None	3
12	Which of the following tasks were not supposed to be performed by IBM Watson during the U.S. Television game show in 2011	Deconstruct the provided clue into words and phrases	Use Hadoop to coordinate the search for a response across terabytes of data	Actuate the buzzer	Provide a semantically correct response in English	4

(Mrs. Nilam Sachin Patil)

Subject Teacher

DA---Unit I

This sheet is for 1 Mark questions							
S.r No	Question	Image	a	b	c	d	Correct Answer
e.g 1	Write down question	img.jpg	Option a	Option b	Option c	Option d	a/b/c/d
1	Business intelligence (BI) is a broad category of application programs which includes _____		a) Decision support b) Data mining c) OLAP d) All of the mentioned				d
2	BI can catalyze a business's success in terms of _____		a) Distinguish the products and services that drive revenues b) Rank customers and locations based on profitability		c) Ranks customers and locations based on probability d) All of the mentioned		d
3	Which of the following areas are affected by BI?		a) Revenue b) CRM c) Sales d) All of the mentioned				b
4	_____ is a performance management tool that recapitulates an organization's performance from several standpoints on a single page.		a) Balanced Scorecard b) Data Cube c) Dashboard d) All of the mentioned				a
5	_____ is a system where operations like data extraction, transformation and loading operations are executed.		a) Data staging b) Data integration c) ETL d) None of the mentioned				a
6	_____ is a category of applications and technologies for presenting and analyzing corporate and external data.		a) Data warehouse b) MIS c) EIS d) All of the mentioned				c
7	Which of the following is the process of basing an organization's actions and decisions on actual measured results of performance?		a) Institutional performance management b) Gap analysis c) Slice and Dice d) None of the mentioned				a
8	Which of the following does not form part of BI Stack in SQL Server?		a) SSRS b) SSIS c) SSAS d) OBIEE				d
9	BI can catalyze a business's success in terms of _____		a) Distinguish the products and services that drive revenues b) Rank customers and locations based on profitability c) Ranks customers and locations based on probability d) All of the mentioned				d
10	This is an approach to selling goods and services in which a prospect explicitly agrees in advance to receive marketing information.		A. customer managed relationship B. data mining C. permission marketing D. one-to-one marketing				c
11	In an Internet context, this is the practice of tailoring Web pages to individual users' characteristics or preferences.		a. Web services b. customer-facing c. client/server d. personalization				d
12	This is the processing of data about customers and their relationship with the enterprise in order to improve the enterprise's future sales and service and lower cost.		a. clickstream analysis b. database marketing c. customer relationship management d. CRM analytics				d
13	This is a broad category of applications and technologies for gathering, storing, analyzing, and providing access to data to help enterprise users make better business decisions.		a. best practice b. data mart c. business information warehouse d. business intelligence				d
14	This is a systematic approach to the gathering, consolidation, and processing of consumer data (both for customers and potential customers) that is maintained in a company's databases.		a. database marketing b. marketing encyclopedia c. application integration d. service oriented integration				a
15	This is an arrangement in which a company outsources some or all of its customer relationship management functions to an application service provider (ASP).		a. spend management b. supplier relationship management c. hosted CRM d. Customer Information Control System				c

16	This is an XML-based metalanguage developed by the Business Process Management Initiative (BPMI) as a means of modeling business processes, much as XML is, itself, a metalanguage with the ability to model enterprise data.		a. BizTalk	b. BPML	c. e-biz	d. ebXML	b
17	This is a central point in an enterprise from which all customer contacts are managed.		a. contact center	b. help system	c. multichannel marketing	d. call center	a
18	This is the practice of dividing a customer base into groups of individuals that are similar in specific ways relevant to marketing, such as age, gender, interests, spending habits, and so on.		a. customer service chat	b. customer managed relationship	c. customer life cycle	d. customer segmentation	d
19	In data mining, this is a technique used to predict future behavior and anticipate the consequences of change.		a. predictive technology	b. disaster recovery	c. phase change	d. predictive modeling	d
20	1. According to analysts, for what can traditional IT systems provide a foundation when they're integrated with big data technologies like Hadoop?		Big data management and data mining	Data warehousing and business intelligence	Management of Hadoop clusters	Collecting and storing unstructured data	a
21	All of the following accurately describe Hadoop, EXCEPT:	Open source	Real-time	Java-based	Distributed computing approach	b	
22	has the world's largest Hadoop cluster.	Apple	Datamatics	Facebook	None of the mentioned	c	
23	What are the five V's of Big Data?	Volume	velocity	Variety	All of the above	d	
24	hides the limitations of Java behind a powerful and concise Clojure API for Cascading.	Scalding	Cascalog	Hcatalog	Hcatalog	b	
25	What are the main components of Big Data?	MapReduce	HDFS	YARN	All of these	d	
26	What are the different features of Big Data Analytics?	Open-Source	Scalability	Data Recovery	All the above	d	
27	Define the Port Numbers for NameNode, Task Tracker and Job Tracker.	NameNode	Task Tracker	Job Tracker	All of the above	d	
28	Facebook Tackles Big Data With based on Hadoop	Project Prism	Prism	ProjectData	ProjectBid	a	
29	What is a unit of data that flows through a Flume agent?	Record	Event	Row	Log	b	
30	A feature F1 can take certain value: A, B, C, D, E, & F and represents grade of students from a college. Which of the following statement is true in the following case	Feature F1 is an example of nominal variable.	Feature F1 is an example of ordinal variable.	It doesn't belong to any of the above category.	Both of these	b	
31	Which of the following is an example of a deterministic algorithm?	PCA	K-Means	None of the above	all of the above	a	
32	What is the entropy of the target variable?	$-(5/8 \log(5/8) + 3/8 \log(3/8))$	$5/8 \log(5/8) + 3/8 \log(3/8)$	$5/8 \log(5/8) + 3/8 \log(3/8)$	$5/8 \log(5/8) - 3/8 \log(5/8)$	a	
33	Point out the correct statement.	a) OLAP is an umbrella term that refers to an assortment of software applications for analyzing an organization's raw data for intelligent decision making	b) Business intelligence equips enterprises to gain business advantage from data	c) BI makes an organization agile thereby giving it a lower edge in today's evolving market condition	None of the mentioned	b	
34	BI can catalyze a business's success in terms of	a) Distinguish the products and services that drive revenues	b) Rank customers and locations based on profitability	c) Ranks customers and locations based on probability	d) All of the mentioned	d	
35	Which of the following areas are affected by BI?	a) Revenue	b) CRM	c) Sales	d) All of the mentioned	b	
36	Which of the following does not form part of BI Stack in SQL Server?	a) SSRS	b) SSIS	c) SSAS	d) OBIEE	d	
37	BI can catalyze a business's success in terms of	a) Distinguish the products and services that drive revenues	b) Rank customers and locations based on profitability	c) Ranks customers and locations based on probability	d) All of the mentioned	d	
38	Heuristic is	A set of databases from different vendors, possibly using different database paradigms	An approach to a problem that is not guaranteed to work but performs well in most cases	Information that is hidden in a database and that cannot be recovered by a simple SQL query.	None of these	b	
39	In an Internet context, this is the practice of tailoring Web pages to individual users' characteristics or preferences.	a. Web services	b. customer-facing	c. client/server	d. personalization	d	
40	Heterogeneous databases referred to	A set of databases from different b vendors, possibly using different database paradigms	An approach to a problem that is not guaranteed to work but performs well in most cases.	Information that is hidden in a database and that cannot be recovered by a simple SQL query.	None of these	a	

DA---Unit II

This sheet is for 1 Mark questions							
S.r No	Question	Image	a	b	c	d	Correct Answer
e.g 1	Write down question	img.jpg	Option a	Option b	Option c	Option d	a/b/c/d
1	Movie Recommendation systems are an example of:		Classification	Clustering	Reinforcement Learning	Regression	b and c
2	Sentiment Analysis is an example of:		Regression	Classification	Clustering	Reinforcement Learning	a,b and d
3	What is the minimum no. of variables/ features required to perform clustering?		0	1	2	3	b
4	Is it possible that Assignment of observations to clusters does not change between successive iterations in K-Means		Yes	No	Can't say	None of these	a
5	Which of the following can act as possible termination conditions in K-Means?		For a fixed number of iterations.	Assignment of observations to clusters does not change between iterations. Except for cases with a bad local minimum.	Centroids do not change between successive iterations.	Terminate when RSS falls below a threshold.	a,b,c,d
6	Which of the following clustering algorithms suffers from the problem of convergence at local optima?		K- Means clustering algorithm	Agglomerative clustering algorithm	Expectation-Maximization clustering algorithm	Diverse clustering algorithm	a and c
7	Which of the following algorithm is most sensitive to outliers?		K-means clustering algorithm	K-medians clustering algorithm	K-modes clustering algorithm	K-medoids clustering algorithm	a
8	How can Clustering (Unsupervised Learning) be used to improve the accuracy of Linear Regression model (Supervised Learning):		Creating different models for different cluster groups.	Creating an input feature for cluster ids as an ordinal variable.	Creating an input feature for cluster centroids as a continuous variable.	Creating an input feature for cluster size as a continuous variable.	a,b,c,d
9	What could be the possible reason(s) for producing two different dendograms using agglomerative clustering algorithm for the same dataset?		Proximity function used	of data points used	of variables used	All of the above	d
10	In which of the following cases will K-Means clustering fail to give good results?		Data points with outliers	Data points with different densities	Data points with round shapes	Data points with non-convex shapes	a,b, and d
11	Which of the following is/are valid iterative strategy for treating missing values before clustering analysis?		Imputation with mean	Nearest Neighbor assignment	Imputation with Expectation Maximization algorithm	All of the above	c
12	Feature scaling is an important step before applying K-Mean algorithm. What is reason behind this?		In distance calculation it will give the same weights for all features	You always get the same clusters. If you use or don't use feature scaling	In Manhattan distance it is an important step but in Euclidian it is not	None of these	a
13	Which of the following method is used for finding optimal of cluster in K-Mean algorithm?		Elbow method	Manhattan method	Euclidian mehtod	All of the above	a
14	What is true about K-Mean Clustering?		K-means is extremely sensitive to cluster center initializations	Bad initialization can lead to Poor convergence speed	Bad initialization can lead to bad overall clustering	None of these	d
15	Which of the following can be applied to get good results for K-means algorithm corresponding to global minima?		Try to run algorithm for different centroid initialization	Adjust number of iterations	Find out the optimal number of clusters	None of these	a,b,c
16	If you are using Multinomial mixture models with the expectation-maximization algorithm for clustering a set of data points into two clusters, which of the assumptions are important:		All the data points follow two Gaussian distribution	All the data points follow n Gaussian distribution (n >2)	All the data points follow two multinomial distribution	All the data points follow n multinomial distribution (n >2)	c

17	Which of the following is/are not true about Centroid based K-Means clustering algorithm and Distribution based expectation-maximization clustering algorithm:		Both starts with random initializations	Both are iterative algorithms	Both have strong assumptions that the data points must fulfill	Expectation maximization algorithm is a special case of K-Means	d
18	Which of the following is/are not true about DBSCAN clustering algorithm:		For data points to be in a cluster, they must be in a distance threshold to a core point	It has strong assumptions for the distribution of data points in dataspace	It has substantially high time complexity of order O(n3)	It does not require prior knowledge of the no. of desired clusters	b and c
19	Which of the following are the high and low bounds for the existence of F-Score?	[0,1]	(0,1)	[-1,1]		None of the above	a
20	1. All of the following increase the width of a confidence interval except:	a. Increased confidence level b. Increased variability c. Increased sample size d. Decreased sample size					c
21	3The p-value in hypothesis testing represents which of the following: Please select the best answer of those provided below.	a. The probability of failing to reject the null hypothesis, given the observed results b. The probability that the null hypothesis is true, given the observed results		c. The probability that the observed results are statistically significant, given that the null hypothesis is true d. The probability of observing results as extreme or more extreme than currently observed, given that the null hypothesis is true			d
22	4. Assume that the difference between the observed, paired sample values is defined in the same manner and that the specified significance level is the same for both hypothesis tests. Using the same data, the statement that "a paired/dependent two sample t-test is equivalent to a one sample t-test on the paired differences, resulting in the same test statistic, same p-value, and same conclusion" is: Please select the best answer of those provided below.	a. Always True b. Never True c. Sometimes True d. Not Enough Information					a
23	19. Green sea turtles have normally distributed weights, measured in kilograms, with a mean of 134.5 and a variance of 49.0. A particular green sea turtle's weight has a z-score of -2.4. What is the weight of this green sea turtle? Round to the nearest whole number.	a. 17 kg b. 151 kg c. 118 kg d. 252 kg					c
24	What percentage of measurements in a dataset fall above the median?	a. 49% b. 50% c. 51% d. Cannot Be Determined					d
25	24. The proportion of variation in 5k race times that can be explained by the variation in the age of competitive male runners was approximately 0.663. What is the value of the sample linear correlation coefficient? Round to 3 decimal places.	a. 0.663 b. 0.814 c. -0.814 d. 0.440					c
26	25. Using all of the results provided, is it reasonable to predict the 5k race time (minutes) of a competitive male runner 73 years of age?	a. Yes; linear correlation between age and 5k race times is statistically significant b. Yes; both the sample linear regression equation and an age in years is provided c. No; linear correlation between age and 5k race times is not statistically significant d. No; the age provided is beyond the scope of our available sample data					d
27	Algorithm is	It uses machine-learning techniques. Here program can learn from past experience and adapt themselves to new situations	Computational procedure that takes some value as input and produces some value as output	Science of making machines performs tasks that would require intelligence when performed by humans	None of these		b

	Bias is		A class of learning algorithm that tries to find an optimum classification of a set of examples using the probabilistic theory	Any mechanism employed by a learning system to constrain the search space of a hypothesis	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.	None of these	
28							b
	Classification is		A subdivision of a set of examples into a number of classes	A measure of the accuracy, of the classification of a concept that is given by a certain theory	The task of assigning a classification to a set of examples	None of these	
29							a
	Binary attribute are		This takes only two values. In general, these values will be 0 and 1 and .they can be coded as one bit	The natural environment of a certain species	Systems that can be used without knowledge of internal operations	None of these	
30							a
	Classification accuracy is		A subdivision of a set of examples into a number of classes	Measure of the accuracy, of the classification of a concept that is given by a certain theory	The task of assigning a classification to a set of examples	None of these	
31							b
	Cluster is		Group of similar objects that differ significantly from other objects	Operations on a database to transform or simplify data in order to prepare it for a machine-learning algorithm	Symbolic representation of facts or ideas from which information can potentially be extracted	None of these	
32							a
	A definition of a concept is----if it recognizes all the instances of that concept		Complete	Consistent	Constant	None of these	
33							a
	A definition or a concept is----- if it classifies any examples as coming within the concept		Complete	Consistent	Constant	None of these	
34							b
	Data selection is		The actual discovery phase of a knowledge discovery process	The stage of selecting the right data for a KDD process	A subject-oriented integrated time variant non-volatile collection of data in support of management	None of these	
35							b
	Classification task referred to		A subdivision of a set of examples into a number of classes	A measure of the accuracy, of the classification of a concept that is given by a certain theory	The task of assigning a classification to a set of examples	None of these	
36							c
	Hybrid is		Combining different types of method or information	Approach to the design of learning algorithms that is structured along the lines of the theory of evolution.	Decision support systems that contain an information base filled with the knowledge of an expert formulated in terms of if-then rules.	None of these	
37							a
	Discovery is		It is hidden within a database and can only be recovered if one is given certain clues (an example IS encrypted information).	The process of executing implicit previously unknown and potentially useful information from data	An extremely complex molecule that occurs in human chromosomes and that carries genetic information in the form of genes.	None of these	
38							b

39	What could be the possible reason(s) for producing two different dendograms using agglomerative clustering algorithm for the same dataset?		Proximity function used	of data points used	of variables used	All of the above	d
40	Is it possible that Assignment of observations to clusters does not change between successive iterations in K-Means		Yes	No	Can't say	None of these	a

DA--Unit III

This sheet is for 1 Mark questions							
S.r No	Question	Image	a	b	c	d	Correct Answer
e.g 1	Write down question	img.jpg	Option a	Option b	Option c	Option d	a/b/c/d
1	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	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	a
2	The correlation coefficient for two real-valued attributes is -0.85 . What does this value tell you?		The attributes are not linearly related.	As the value of one attribute decreases the value of the second attribute increases.	As the value of one attribute increases the value of the second attribute also increases.	The attributes show a linear relationship	b
3	Given a rule of the form IF X THEN Y, rule confidence is defined as the conditional probability that		Y is false when X is known to be false.	Y is true when X is known to be true.	X is true when Y is known to be true	X is false when Y is known to be false.	b
4	Chameleon is		Density based clustering algorithm	Partitioning based algorithm	Model based algorithm	Hierarchical clustering algorithm	d
5	Find odd man out		DBSCAN	K-Mean	PAM	None of above	a

6	The number of iterations in apriori		increases with the size of the data	decreases with the increase in size of the data	increases with the size of the maximum frequent set	decreases with increase in size of the maximum frequent set	c
7	Which of the following are interestingness measures for association rules?		Recall	Lift	Accuracy	All of Above	b
8	Given a frequent itemset L, If $ L = k$, then there are		$2^k - 2$ candidate association rules	2^k candidate association rules	2^{k-2} candidate association rules	2^{k-2} candidate association rules	c
9	_____ is an example for case based-learning		Decision trees	Neural networks	Genetic algorithm	K-nearest neighbor	d
10	The average positive difference between computed and desired outcome values.		mean positive error	mean squared error	mean absolute error	root mean squared error	c
11	Frequent item sets is		Superset of only closed frequent item sets	Superset of only maximal frequent item sets	Subset of maximal frequent item sets	Superset of both closed frequent item sets and maximal frequent item sets	d
12	Assume that we have a dataset containing information about 200 individuals. A supervised data mining session has discovered the following rule: IF age < 30 & credit card insurance = yes THEN life insurance = yes Rule Accuracy: 70% and Rule Coverage: 63% How many individuals in the class life insurance= no have credit card insurance and are less than 30 years old?			63	38	40	89
13	Which of the following is cluster analysis?		Simple segmentation	Grouping similar objects	Labeled classification	Query results grouping	b
14	A good clustering method will produce high quality clusters with		high inter class similarity	high intra class similarity	low intra class similarity	None of above	c
15	Which two parameters are needed for DBSCAN		Min threshold	Min points and eps	Min sup and min confidence	Number of centroids	b

16	Which statement is true about neural network and linear regression models?		Both techniques build models whose output is determined by a linear sum of weighted input attribute values.	The output of both models is a categorical attribute value.	Both models require numeric attributes to range between 0 and 1.	Both models require input attributes to be numeric.	d
17	In Apriori algorithm, if 1 item-sets are 100, then the number of candidate 2 item-sets are		100	200	4950	5000	c
18	Significant Bottleneck in the Apriori algorithm is		Finding frequent itemsets	Pruning	Candidate generation	Number of iterations	c
19	Machine learning techniques differ from statistical techniques in that machine learning methods		are better able to deal with missing and noisy data	typically assume an underlying distribution for the data	have trouble with large-sized datasets	are not able to explain their behavior.	a
20	The probability of a hypothesis before the presentation of evidence.		a priori	posterior	conditional	subjective	a
21	KDD represents extraction of		data	knowledge	rules	model	b
22	Which statement about outliers is true?		Outliers should be part of the training dataset but should not be present in the test data.	Outliers should be identified and removed from a dataset.	The nature of the problem determines how outliers are used	Outliers should be part of the test dataset but should not be present in the training data.	c
23	The most general form of distance is		Manhattan	Eucledian	Mean	Minkowski	d
24	Which Association Rule would you prefer		High support and medium confidence	High support and low confidence	Low support and high confidence	Low support and low confidence	c
25	In a Rule based classifier, If there is a rule for each combination of attribute values, what do you call that rule set R		Exhaustive	Inclusive	Comprehensive	Mutually exclusive	a
26	The apriori property means		If a set cannot pass a test, its supersets will also fail the same test	To decrease the efficiency, do level-wise generation of frequent item sets	To improve the efficiency, do level-wise generation of frequent item sets	If a set can pass a test, its supersets will fail the same test	a

27	If an item set ‘XYZ’ is a frequent item set, then all subsets of that frequent item set are		Undefined	Not frequent	Frequent	Can not say	c
28	The probability that a person owns a sports car given that they subscribe to automotive magazine is 40%. We also know that 3% of the adult population subscribes to automotive magazine. The probability of a person owning a sports car given that they don't subscribe to automotive magazine is 30%. Use this information to compute the probability that a person subscribes to automotive magazine given that they own a sports car		0.0368	0.0396	0.0389	0.0398	b
29	Simple regression assumes a _____ relationship between the input attribute and output attribute.		quadratic	inverse	linear	reciprocal	c
30	To determine association rules from frequent item sets		Only minimum confidence needed	Neither support nor confidence needed	Both minimum support and confidence needed	Minimum support is needed	c
31	If {A,B,C,D} is a frequent itemset, candidate rules which is not possible is		C → A	D → ABCD	A → BC	B → ADC	b
32	Which Association Rule would you prefer		High support and low confidence	Low support and high confidence	Low support and low confidence	High support and medium confidence	b
33	Classification rules are extracted from		decision tree	root node	branches	siblings	a
34	What does K refers in the K-Means algorithm which is a non-hierarchical clustering approach?		Complexity	Fixed value	No of iterations	. number of clusters	d
35	If Linear regression model perfectly fit i.e., train error is zero, then		Test error is also always zero	Test error is non zero	Couldn't comment on Test error	Test error is equal to Train error	c
	Which of the following metrics can be used for evaluating regression models? i) R Squared ii) Adjusted R Squared iii) F Statistics iv) RMSE/MSE/MAE		ii and iv	i and ii	ii, iii and iv	i, ii, iii and iv	d

37	How many coefficients do you need to estimate in a simple linear regression model (One independent variable)?		1	2	3	4	b
38	In a simple linear regression model (One independent variable), If we change the input variable by 1 unit. How much output variable will change?		by 1	no change	by intercept	by its slope	d
39	In syntax of linear model lm(formula,data,,..), data refers to _____		Matrix	array	vector	list	c
40	In the mathematical Equation of Linear Regression $Y = \beta_1 + \beta_2 X + \epsilon$, (β_1, β_2) refers to _____		(X-intercept, Slope)	(Slope, X-Intercept)	(Y-Intercept, Slope)	(slope, Y-Intercept)	c

DA---Unit IV

This sheet is for 1 Mark questions								
S.r No	Question	Image	a	b	c	d	Correct Answer	
e.g 1	Write down question	img.jpg	Option a	Option b	Option c	Option d	a/b/c/d	
1	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	a	
3	What is Decision Tree?	Flow-Chart		which internal node represents test on an attribute, each	Structure in which internal node represents test	None of Above	c	
4	Decision Trees can be used for Classification Tasks.		TRUE	FALSE			a	
5	Choose from the following that are Decision Tree nodes?		Decision Nodes	End Nodes	Chance Nodes	All of Above	d	
6	Decision Nodes are represented by _____		Disks	Squares	Circles	Triangles	b	
7	Chance Nodes are represented by _____		Disks	Squares	Circles	Triangles	c	
8	End Nodes are represented by _____		Disks	Squares	Circles	Triangles	d	
9	Which of the following are the advantage/s of Decision Trees?		Possible Scenarios can be added	Use a white box model, If given result is provided by a model	Worst, best and expected values can be determined for different scenarios	All of Above	d	

10	Which of the following statements about Naive Bayes is incorrect?		Attributes are equally important.	Attributes are statistically dependent of one another given the class value.	Attributes are statistically independent of one another given the class value.	Attributes can be nominal or numeric	b
11	Which of the following is not supervised learning?		Clustering	Decision Tree	Linear Regression	Naive Bayesian	a
12	How many terms are required for building a bayes model?			1	2	3	4 c
13	Where does the bayes rule can be used?		Solving queries	Increasing complexity	Decreasing complexity	Answering probabilistic query	d
14	How the bayesian network can be used to answer any query?		Full distribution	Joint distribution	Partial distribution	All of Above	b
15	What is the consequence between a node and its predecessors while creating bayesian network?		Functionally dependent	Dependant	Conditionally independent	Both Conditionally dependant & Dependant	c

16	Bayesian classifiers is	A class of learning algorithm that tries to find an optimum classification of a set of examples using the probabilistic theory.	Any mechanism employed by a learning system to constrain the search space of a hypothesis	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.	None of these	a

17	Bias is		A class of learning algorithm that tries to find an optimum classification of a set of examples using the probabilistic theory	Any mechanism employed by a learning system to constrain the search space of a hypothesis	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.	None of these	b
18	Background knowledge referred to		Additional acquaintance used by a learning algorithm to facilitate the learning process	A neural network that makes use of a hidden layer	It is a form of automatic learning.	None of these	a
19	Classification accuracy is		A subdivision of a set of examples into a number of classes	A measure of the accuracy, of the classification of a concept that is given by a certain theory	The task of assigning a classification to a set of examples	None of these	b

20	Classification is		A subdivision of a set of examples into a number of classes	A measure of the accuracy, of the classification of a concept that is given by a certain theory	The task of assigning a classification to a set of examples	None of these	a
21	Discovery is		It is hidden within a database and can only be recovered if one is given certain clues (an example IS encrypted information).	The process of executing implicit previously unknown and potentially useful information from data	An extremely complex molecule that occurs in human chromosomes and that carries genetic information in the form of genes.	None of these	b
22	Classification task referred to		A subdivision of a set of examples into a number of classes	A measure of the accuracy, of the classification of a concept that is given by a certain theory	The task of assigning a classification to a set of examples	None of these	c

23	Euclidean distance measure is		A stage of the KDD process in which new data is added to the existing selection.	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	The distance between two points as calculated using the Pythagoras theorem	None of these	c
24	The problem of finding hidden structure in unlabeled data is called		Supervised learning	Unsupervised learning	Reinforcement learning	None of these	b
25	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		Classification	Regression	Clustering	Structural equation modeling	b
26	Discriminating between spam and ham e-mails is a classification task, true or false?		TRUE	FALSE			a
27	which of the following is not involve in data mining?		Knowledge extraction	Data archaeology	Data exploration	Data transformation	d
28	Naive prediction is		A class of learning algorithms that try to derive a Prolog program from examples	A table with n independent attributes can be seen as an n-dimensional space.	A prediction made using an extremely simple method, such as always predicting the same output.	None of these	c

29	Node is		A component of a network	In the context of KDD and data mining, this refers to random errors in a database table.	One of the defining aspects of a data warehouse	None of these	a
30	Prediction is		The result of the application of a theory or a rule in a specific case	One of several possible entries within a database table that is chosen by the designer as the primary means of accessing the data in the table.	Discipline in statistics that studies ways to find the most interesting projections of multi-dimensional spaces.	None of these	a
31	What is the relation between the distance between clusters and the corresponding class discriminability?		proportional	inversely-proportional	no-relation	None of these	a
32	the classification method in which the upper limit of interval is same as of lower class interval is called....		exclusive method	inclusive method	mid point method	None of these	a
33	larger value is 60 and the smallest value is 40 and the number of classes is 5 then the class interval is		20	25	4	15	c
34	summary and presentation of data in tabular form with several non overlapping classes is referred as		nominal distribution	frequency distribution	ordinal distribution	None of these	b
35	the classification method in which the upper and lower limit of interval is also in class interval itself is called....		exclusive method	inclusive method	mid point method	None of these	b

	Suppose there are 25 base classifiers. Each classifier has error rates of $e = 0.35$. Suppose you are using averaging as ensemble of above 25 classifiers will make a wrong prediction? Note: all classifiers are independent of each other			0.05	0.06	0.07	0.08	b
36	The most widely used metrics and tools to assess a classification model are:		Confusion matrix	Cost-sensitive accuracy	Area under the ROC curve	All of Above	d	
37	When performing regression or classification, which of the following is the correct way to preprocess the data?		Normalize the data → PCA → training	PCA → normalize PCA output → training	Normalize the data → PCA → normalize PCA output → training	None of these	a	
38	Which of the following is true about Naive Bayes ?		Assumes that all the features in a dataset are equally important	Assumes that all the features in a dataset are independent	both a and b	None of these	c	
39	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		1 and 2	2 and 3	1, 2, and 3	1 and 3	c	

S.r No	Question	Image	a	b	c	d	Correct Answer
1	Data visualization is realted with...		Pictorial representaions	numerical representation	numerical calculations	None of these	a
2	Which of the following are Use of data visualization		See context of data	Clear data understanding	finding pattern in data	all of above	d
3	Which of the following statements are true about using visualizations to display a dataset? I. Visualizations are visually appealing, but don't help the viewer understand relationships that exist in the data II. Visualizations like graphs, charts, or visualizations with pictures are useful for conveying information, while tables just filled with text are not useful. III. Patterns that exist in the data can be found more easily by using a visualization		I AND II	II AND III	I AND III	ONLY III	d
4	The plot method on Series and DataFrame is just a simple wrapper around _____		gplt.plot()	plt.plot()	plt.plotgraph()	none of the mentioned	b
5	Point out the correct combination with regards to kind keyword for graph plotting.		'hist' for histogram	'box' for boxplot	'area' for area plots	all of the mentioned	d
6	Which of the following value is provided by kind keyword for barplot?		bar	bar	bar	none of the mentioned	a

7	You can create a scatter plot matrix using the _____ method in pandas.tools.plotting.		sca_matrix	scatter_matrix	DataFrame.plot	all of the mentioned	b
8	Plots may also be adorned with error bars or tables.	True	FALSE	Cannot Tell	All Above	a	
9	Which of the following plots are often used for checking randomness in time series?	Autocausation	Autorank	Autocorrelation	none of the mentioned	c	
10	_____ plots are used to visually assess the uncertainty of a statistic	Lag	RadViz	Bootstrap	All Above	c	
11	Which of the following is not a challenge in Big Data Visualization>?	Velocity	Volume	Version	Variety	c	
12	Which of the following is not a problem in Big Data Visualization>?	Visual Noise	Scaled Data	Large image perception	Information Loss	b	
13	Which of the following is a problem in Big Data Visualization>?	Structured Data	Scaled Data	Visual Noise	Multiple valued Data	c	
14	Which of the candidate is suitable for interactive visualtization?	Type of Visual	Cardinality	Size of data	all of above	d	
15	Which of the following follows interactive visualization approach?	Zoom+Pan	Focus+Context	Overview+Details	all of above	d	
16	Visual Mapping is important for_____	Remapping	Overview+Details	Focus	Context	a	
17	Data visualtization techniques are:	Scatter Plot	Line Chart	Pie Chart	all of above	d	
18	Information Visultization techniques are	Flow Chart	Time Line	DFD	All of above	d	
19	Data visualtization techniques are:	Flow Chart	Time Line	Pie Chart	None of these	c	
20	Information Visultization techniques are	Flow Chart	Line Chart	Pie Chart	None of these	a	
21	Data visualtization techniques are:	Scatter Plot	Time Line	DFD	None of these	a	
22	Information Visultization techniques are	Scatter Plot	Time Line	Bubble Chart	None of these	b	
23	Data visualtization techniques are:	Histogram	Parallel Coordinates	Time Line	None of these	a	

	Information Visualization techniques are		Semantic Network	Histogram	Area Chart	None of these	a
24	Which of the following is realted term with correlation?		Exponential	U-Shape	Null	All of above	d
25	Data visualtization techniques are:		Scatter Plot	Time Line	DFD	None of these	a
26	Coulmn graph is another name for _____		Bar Chart	Scatterplot	Histogram	Area Chart	a
27	Which of the following follows interactive visualization approach?		Zoom+Pan	Focus+Context	Overview+Details	all of above	d
28	information Visualtization techniques are		Pie Chart	Scatterplot	Histogram	Area Chart	a
29	Which of the following is category of timeline?		Linear Timeline	Modular Timeline	Variant Timeline	ER Timeline	a
30	Which of the following specifies relationship amongst variables?		Scatter Plot	Line Chart	Area Chart	All of above	d
31	Which of the following specifies category Proportions?		Pie Chart	Histogram	Bar chart	All of above	d
32	Which of the following is category of timeline?		Variant Timeline	ER Timeline	Comarative Timeline	Modular Timeline	c
33	Information Visualtization techniques are		Flow Chart	Time Line	DFD	All of above	d
34	Data visualtization techniques are:		Flow Chart	Time Line	Pie Chart	None of these	c
35	Data visualtization is realted with...		Pictorial representaions	numerical representation	numerical calculations	None of these	a
36	Which of the following follows interactive visualization approach?		Zoom+Pan	Focus+Context	Overview+Details	all of above	d
37	Which of the following are Use of data visualization		See context of data	Clear data understanding	finding pattern in data	all of above	d
38	Which of the following specifies relationship amongst variables?		Pie Chart	Histogram	Area Chart	None of these	c
39	Which of the following specifies category Proportions?		Pie Chart	Scatter Plot	Line Chart	None of these	a
40							

This sheet is for 1 Mark questions							
S.r No	Question	Image	a	b	c	d	Correct Answer
e.g 1	Write down question	img.jpg	Option a	Option b	Option c	Option d	a/b/c/d
1	Precies and steady format data is_____		Structured Data	Un Structured Data	semi Structured Data	Quasi Structured Data	a
2	Inconsistant Data is_____		Structured Data	Un Structured Data	semi Structured Data	Quasi Structured Data	b
3	Format that self defines itself is_____		Structured Data	Un Structured Data	semi Structured Data	Quasi Structured Data	c
4	A little Bit inconsistant data is_____		Structured Data	Un Structured Data	semi Structured Data	Quasi Structured Data	d
5	XML is an example of_____		Structured Data	Un Structured Data	semi Structured Data	Quasi Structured Data	
6	RDBMS Follows_____		Structured Data	Un Structured Data	semi Structured Data	Quasi Structured Data	a
7	Watson is developed by_____		IBM	Microsoft	AT&T	Google	a
8	Hadoop is _____ based Framework.		C++	Python	JAVA	C#	c
9	Which of the following are components of Hadoop?		MAPREDUCE	YARN	HDFS	All of Above	d
10	Which of the following are components of HIVE?		JDBC	Thrift Server	CLI	All of Above	d
11	Mahout provides_____		JAVA Executable Libraries	C# Executables	Mountable Image Format	All of Above	a
12	Which of the following are components of HIVE?		FLATTEN	Thrift Server	Muster	None of these	b
13	Which of the following are components of HIVE?		FLATTEN	Thrift Server	Muster	All of above	b
14	Which of the following is components of Hadoop?		Fork	YARN	CLI	Metadata	b
15	RDBMS Follows_____		Structured Data	Un Structured Data	semi Structured Data	Quasi Structured Data	a
16	Which of the following is a clustering techique?		Fuzzy K means	Canopy	K-Means	All of above	d
17	Which of the following is HBASE Data Model Terminology?		Row	Table	Column	All of Above	d

18	Which of the following is not a classification technique?		Logistic Regression	Random Forest	Recommender Algo	Naïve Bayes	c
19	Which of the following is a classification technique?		Logistic Regression	Random Forest	Naïve Bayes	All of Above	d
20	Which of the following is HBASE Data Model Terminology?		Column Family	Cell	Timestamp	All of Above	d
21	Which of the following is a clustering technique?		Logistic Regression	Random Forest	K-Means	Naïve Bayes	c
22	Which of the following is HBASE Data Model Terminology?		Identifier	Variant	Timestamp	None of the above	c
23	Which of the following is not a classification technique?		Logistic Regression	Random Forest	K-Means	Naïve Bayes	c
24	Which of the following are components of HIVE?		FLATTEN	Thrift Server	Muster	None of these	b
25	Which of the following is HBASE Data Model Terminology?		Identifier	Variant	Column Qualifier	None of the above	c
26	Mahout provides_____		JAVA Executable Libraries	C# Executables	Mountable Image Format	None of the above	a
27	Which of the following is not a clustering technique?		Logistic Regression	Canopy	K-Means	Fuzzy K means	a
28	Which of the following is a clustering technique?		Fuzzy K means	Canopy	K-Means	All of above	d
29	Point out the correct statement.		Hadoop do need specialized hardware to process the data	Hadoop 2.0 allows live stream processing of real-time data	In Hadoop programming framework output files are divided into lines or records	None of the above	b
30							
31	What was Hadoop named after?		Creator Doug Cutting's favorite circus act	Cutting's high school rock band	The toy elephant of Cutting's son	A sound Cutting's laptop made during Hadoop development	c

	programming model used to develop Hadoop-based applications that can process massive amounts of data.		MapReduce	Mahout	Oozie	None of the above	a
32							
33	Which of the following is not a classification technique?		Logistic Regression	Random Forest	K-Means	Naïve Bayes	c
34	Which of the following are components of HIVE?		FLATTEN	Thrift Server	Muster	All of above	b
35	Which of the following is components of Hadoop?		Fork	YARN	CLI	None of above	b
36	Hadoop is a framework that works with a variety of related tools. Common cohorts include _____		MapReduce, Hive and HBase	MapReduce, MySQL and Google Apps	MapReduce, Hummer and Iguana	All of above	a
37	NoSQL databases is used mainly for handling large volumes of _____ data.		Structured Data	Un Structured Data	semi Structured Data	Quasi Structured Data	b
38	Which of the following is not a phase of Data Analytics Life Cycle?		Communication	Recall	Data Preparation	Model Planning	b
39	Which of the following is a NoSQL Database Type?		SQL	Document databases	JSON	All of above	b
40	Which of the following is not a NoSQL database		SQL Server	MongoDB	Cassandra	None of the above	a

Q. No.	Question	Option A	Option B	Option C	Option D	Answer
UNIT- I						
1	Data can be _____.	Structured	Unstructured	Semi- structured	All of the above	D
2	There are total _____ phases in data analytics life cycle.	5	6	7	4	B
3	GINA stands for _____.	Global innovation netw	Global intarnal network anal	Global irregular networkanalys	Noneof the above	A
4	Communicate result is the _____ phase of data analytics life cycle.	3	5	6	7	B
5	Which of the following is not source of big data?	Surveillance system	Social media	Stock exchange	Traffic signal data	D
6	Which of the following is not the phase of data analytics life cycle?	Communicate result	Model building	Model Planning	None of the above	D
7	The statement: big data can not be processed using traditional computing is _____.	TRUE	FALSE			A
8	Choose the correct statement: S1: Business intelligence requires structured data S2: Data science works with unstructured data.	S1 is true	S2 is true	Both S1 and S2 are true	Both S1 and S2 are false	C
9	Which of the following is not the application of big data?	Fraud detection	Call center analysis	IT log analytics	Traffic signal analysis	D
10	Which of the following is component of current analytical architecture?	Departmental warehos	Dashboard	Data sources	All of the above	D
UNIT- II						
1	A statement made about a population for testing purpose is called?	a) Statistic	b) Hypothesis	c) Level of Significance	d) Test-Statistic	B
2	If the assumed hypothesis is tested for rejection considering it to be true is called?	a) Null Hypothesis	b) Statistical Hypothesis	c) Simple Hypothesis	d) Composite Hypothesi	A
3	What is the assumption made for performing the hypothesis test with T distribution?	a) the distribution is nor	b) the distribution has more	c) the distribution has a constan	d) the distribution follow	D
4	In a two tailed test when a Null Hypothesis is rejected for a True Alternative Hypothesis then it has _____.	a) Type 1 error	b) Type 2 error	c) No error	d) Many errors	C
5	If a hypothesis is rejected at 0.6 Level of Significance then _____.	a) it will be rejected at	b) it must be rejected at 0.5	c) it may be rejected at 0.5 level	d) it cannot be rejected a	C
6	A Null Hypothesis has Level of Significance 9%. For what values of Level of Significances it will be rejected?	a) 0.99	b) 0.009	c) 0.099	d) 0.9	B
7	Consider a trial of a criminal. If a type 1 error has occurred in thee judgement then which of the following statement is true?	a) a guilty person is set	b) an innocent person is con	c) a guilty person is convicted	d) an innocent person is	B
8	Which of the following is required by K-means clustering?	a) defined distance met	b) number of clusters	c) initial guess as to cluster cent	d) all of the mentioned	D
9	Point out the wrong statement.	a) k-means clustering is	b) k-means clustering aims to	c) k-nearest neighbor is same as	d) none of the mentioned	C
10	Which of the following function is used for k-means clustering?	a) k-means	b) k-mean	c) heatmap	d) none of the mentioned	C
UNIT- III						
1	Associtaion rule is _____ method.	Rule-based machine lea	Deep learning	Natural language processing	Noneof the above	A
2	The statement: Association rules does not consider the order of items is _____.	TRUE	FALSE			A
3	Apriori algorithm is method of _____ mining.	Rare itemset	Frequent itemset	Common itemset	All of the above	B
4	In linear regression dependent variable is _____ in nature.	Discrete	Catagorical	Contineous	None of the above	C
5	Which of the following is not typeof the regression analysis?	Ridge regression	Linear regression	Logistic regression	Multiple regression	D
6	In logistic regression dependent variable is _____ in nature.	Binary	Unary	Ternary	Octal	A
7	For non- linearly seperable data which regression model is suitable?	Logistic regression	Linear regression	Polynomial	Lasso	C
8	The statement: Linear regression is suitable whe contineous variables are contineousor descrete with catogorical data types is _____.	TRUE	FALSE			A
9	The statement: regression analysis is used tomodel the relationship between a dependent variable and one or more independent variables is _____.	TRUE	FALSE			A
10	Candidate generation step can be limitation of the Apriorialgorithm?	Yes	No			A
UNIT- IV						
1	Which of the following statements about Naive Bayes is incorrect?	a) Attributes are equally important.	b) Attributes are statistically dependent of	c) Attributes are statistically independent of one another	d) Attributes can be nominal or numeric	B
2	To which does the local structure is associated?	a) Hybrid	b) Dependant	c) Linear	d) None of the mentione	C
3	Examples of Naïve Bayes Algorithm is/are	(A) Spam filtration	(B) Sentimental analysis	(C) Classifying articles	(D) All of the above	D
4	Naïve Bayes algorithm is based on _____ and used for solving classification problems.	(A) Bayes Theorem	(B) Candidate elimination al	(C) EM algorithm	(D) None of the above	A
5	Types of Naïve Bayes Model:	(A) Gaussian	(B) Multinomial	(C) Bernoulli	(D) All of the above	D
6	Disadvantages of Naïve Bayes Classifier:	(A) Naive Bayes assum	(B) It performs well in Multi	(C) Naive Bayes is one of the fa	(D) It is the most popula	A
7	The benefit of Naïve Bayes:-	(A) Naïve Bayes is one	(B) It is the most popular ch	(C) It can be used for Binary as	(D) All of the above	D
8	How the entries in the full joint probability distribution can be calculated?	a) Using variables	b) Using information	c) Both Using variables & infor	d) None of the mentione	B
9	How the bayesian network can be used to answer any query?	a) Full distribution	b) Joint distribution	c) Partial distribution	d) All of the mentioned	B
10	What is the consequence between a node and its predecessors while creating bayesian network?	a) Functionally depende	b) Dependant	c) Conditionally independent	d) Both Conditionally de	C
UNIT- V						
1	Which of the following is challenge to the data visualization?	Velocity	Volume	Vriety	All of the above	D
2	Volume of data refers to _____.	Speed	Size	Source	None of the above	B
3	Whicohfthe following is technique to visual data representation?	Informationvisualizatio	Concept Visualization	Metaphore visualization	All of the above	D

4	Table is the form of _____ data visualization.	Sturctured	Unstructured	Semi-Structured	None of the above	A
5	Which of the following is type of data visualization?	Histogram	Timeline	Both A and B	None	C
6	The statement: Histograms are used to summarize tiny data sets in the form of tables is _____.	TRUE	FALSE			B
7	Scatter plots are also called as _____.	Histograms	Pie Charts	Heat maps	X-Y plots	C
8	Google chart is the example of _____ visualization.	Multidimensional	Single dimensional	Both A and B	None	A
9	Leaflet is _____ data visualization tool.	Proprietary	Open source	Complex	None	B
10	Veracity of data refers to _____.	Variety of data	Complexity of data	Speed of data	Velocity of data	B

UNIT- VI

1	Running a _____ program involves running mapping tasks on many or all of the nodes in our cluster.	a) MapReduce	b) Map	c) Reducer	d) All of the mentioned	A
2	Input to the _____ is the sorted output of the mappers.	a) Reducer	b) Mapper	c) Shuffle	d) All of the mentioned	A
3	The right number of reduces seems to be _____.	a) 0.90	b) 0.80	c) 0.36	d) 0.95	D
4	Point out the wrong statement.	a) Reducer has 2 primaries	b) Increasing the number of reducers	c) It is legal to set the number of reducers	d) The framework group	A
5	The output of the _____ is not sorted in the Mapreduce framework for Hadoop.	a) Mapper	b) Cascader	c) Scalding	d) None of the mentioned	D
6	Which of the following phases occur simultaneously?	a) Shuffle and Sort	b) Reduce and Sort	c) Shuffle and Map	d) All of the mentioned	A
7	Mapper and Reducer implementations can use the _____ to report progress or just indicate that they are alive.	a) Partitioner	b) OutputCollector	c) Reporter	d) All of the mentioned	C
8	_____ is a generalization of the facility provided by the MapReduce framework to collect data output by the Mapper or the Reducer.	a) Partitioner	b) OutputCollector	c) Reporter	d) All of the mentioned	B
9	_____ is the primary interface for a user to describe a MapReduce job to the Hadoop framework for execution.	a) Map Parameters	b) JobConf	c) MemoryConf	d) None of the mentioned	B
10	Mapper implementations are passed the JobConf for the job via the _____ method.	a) JobConfigure.config	b) JobConfigurable.config	c) JobConfigurable.configurable	d) None of the mentioned	B

Q. No.	Question	Option A	Option B	Option C	Option D	Answer
UNIT- I						
1	Velocity of data refers to _____.	Speed of processing	Size of data	Complexity of data	None of the above	A
2	Variety of data refers to _____.	Different types and forms of data	Sources of data	Both A & B	None of the above	A
3	Rapid growth in data is referred as _____.	Data Conversion	Data diversion	Data explosion	All of the above	C
4	Which among the following is valid source of data?	Clickstream	Contracts	Application log	All of the above	D
5	Which of the following is not source of big data?	Surveillance system	Social media	Stock exchange	Traffic signal data	D
6	Which of the following is not the phase of data analytics life cycle?	Communicate result	Model building	Model Planning	None of the above	D
7	The statement: big data can not be processed using traditional computing is _____.	TRUE	FALSE			A
8	Choose the correct statement: S1: Business intelligence requires structured data S2: Data science works with unstructured data.	S1 is true	S2 is true	Both S1 and S2 are true	Both S1 and S2 are false	C
9	Which of the following is not the application of big data?	Fraud detection	Call center analysis	IT log analytics	Traffic signal analysis	D
10	Which of the following is component of current analytical architecture?	Departmental warehouse	Dashboard	Data sources	All of the above	D
UNIT- II						
1	A statement whose validity is tested on the basis of a sample is called?	a) Null Hypothesis	b) Statistical Hypothesis	c) Simple Hypothesis	d) Composite Hypothesis	B
2	A hypothesis which defines the population distribution is called?	a) Null Hypothesis	b) Statistical Hypothesis	c) Simple Hypothesis	d) Composite Hypothesis	C
3	If the null hypothesis is false then which of the following is accepted?	a) Null Hypothesis	b) Positive Hypothesis	c) Negative Hypothesis	d) Alternative Hypothesis.	D
4	The probability of Type I error is referred as?	a) $1-\alpha$	b) β	c) α	d) $1-\beta$	B
5	Alternative Hypothesis is also called as?	a) Composite hypothesis	b) Research Hypothesis	c) Simple Hypothesis	d) Null Hypothesis	B
6	In a study, subjects are randomly assigned to one of three groups: control, experimental A, or experimental B. After treatment, the mean scores for the three groups are 100, 102, and 104 respectively. The null hypothesis is that the three groups have equal means. The alternative hypothesis is that the three groups do not have equal means. The test statistic is 2.52. The critical value at the 0.05 level of significance is 2.77. Then what is the conclusion?	a. the correlation coefficient	b. chi square	c. the t-test	d. the analysis of variance	D
7	When the k population means are truly different from each other, it is likely that the average error deviation:	a. is relatively large compared to b. is relatively small compared to c. is about equal to the average d. none of the above				B
8	K-means is not deterministic and it also consists of number of iterations.	a) True	b) False			A
9	The _____ sum of squares measures the variability of the sample treatment means around the overall mean.	a. treatment	b. error	c. interaction	d. total	A
10	As variability due to chance decreases, the value of F will	a. increase	b. stay the same	c. decrease	d. can't tell from the given info	A
UNIT- III						
1	If Linear regression model perfectly first i.e., train error is zero, then	a) Test error is also always zero	b) Test error is non zero	c) Couldn't comment on Test	d) Test error is equal to Train	C
2	How many coefficients do you need to estimate in a simple linear regression model (One independent variable)?	a) 1	b) 2	c) 3	d) 4	B
3	In a simple linear regression model (One independent variable), If we change the input variable by 1 unit. How much output variable will change?	a) by 1	b) no change	c) by intercept	d) by its slope	D
4	_____ is a simple approach to supervised learning. It assumes that the dependence of Y on X1, X2, ..., Xp is linear.	a) Linear regression	b) Logistic regression	c) Gradient Descent	d) Greedy algorithms	A
5	When there are more than one independent variables in the model, then the linear model is termed as	a) Unimodal	b) Multiple model	c) Multiple Linear model	d) Multiple Logistic model	C
6	Logistic regression is used when you want to:	A. Predict a dichotomous variable	B. Predict a continuous variable	C. Predict any categorical variable	D. Predict a continuous variable	A
7	Logistic regression assumes a:	A. Linear relationship between co	B. Linear relationship between co	C. Linear relationship between co	D. Linear relationship between co	B
8	In binary logistic regression:	A. The dependent variable is con	B. The dependent variable is divi	C. The dependent variable co	D. There is no dependent varia	C
9	Which of the following is direct application of frequent itemset mining	Market basket analysis	Intrusion detection	Outlier detection	Social network analysis	A
10	What do you mean by support(A)?	Total number of transactions	Total Number of transactions	Number of transactions co	Number of transactions not	C
UNIT- IV						
1	End Nodes are represented by _____.	a) Disks	b) Squares	c) Circles	d) Triangles	D
2	Which of the following are the advantage/s of Decision Trees?	a) Possible Scenarios can be addde	b) Use a white box model, If give	c) Worst, best and expected v	d) All of the mentioned	D
3	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,	a) Decision tree	b) Graphs	c) Trees	d) Neural Networks	A
4	How the decision tree reaches its decision?	a) Single test	b) Two test	c) Sequence of test	d) No test	C
5	How many terms are required for building a bayes model?	a) 1	b) 2	c) 3	d) 4	C
6	Where does the bayes rule can be used?	a) Solving queries	b) Increasing complexity	c) Decreasing complexity	d) Answering probabilistic que	D
7	What does the bayesian network provides?	a) Complete description of the do	b) Partial description of the doma	c) Complete description of th	d) None of the mentioned	A
8	How the entries in the full joint probability distribution can be calculated?	a) Using variables	b) Using information	c) Both Using variables & inf	d) None of the mentioned	B
9	How the bayesian network can be used to answer any query?	a) Full distribution	b) Joint distribution	c) Partial distribution	d) All of the mentioned	B
10	What is the consequence between a node and its predecessors while creating bayesian network?	a) Functionally dependent	b) Dependant	c) Conditionally independent	d) Both Conditionally dependa	C
UNIT- V						
1	Variety of data refers to _____.	Different types and forms of data	Various types of data	Both A & B	None of the above	C
2	Velocity of data refers to _____.	Processing speed of data	Volume of data	Variety of data	None of the above	A
3	Visual analytics has _____ challenges.	Scalability	Dynamics	Only A	Both A & B	D
4	Metadata means _____.	Data about data	Huge data	Tiny data	None of the above	A
5	Data visualization may be challenging if the size and dimension of data _____.	Huge	Small	Moderate	None of the above	A
6	Data visualization may be challenging due to _____.	Heterogeneity	Diversity	Both A & B	None of the above	C
7	Most preferred factor in big data analysis is _____.	Analysis speed	Size of data	Tools used	None of the above	A
8	Zoom in and zoom out are the types of _____ data visualization approach.	Interactive	Communicative	Transfer	All of the above	A
9	_____ is the method of data visualization to visualize plans.	Strategic visualization	Metaphor Visualization	Concept Visualization	Information Visualization	C
10	Different graphic representation formats are merged in _____.	Strategic visualization	Metaphor Visualization	Concept Visualization	Compound Visualization	D
UNIT- VI						
1	A _____ node acts as the Slave and is responsible for executing a Task assigned to it by the JobTracker.	a) MapReduce	b) Mapper	c) TaskTracker	d) JobTracker	C
2	Point out the correct statement.	a) MapReduce tries to place the d	b) Map Task in MapReduce is p	c) Reduce Task in MapReduc	d) All of the mentioned	A

3	part of the MapReduce is responsible for processing one or more chunks of data and producing the output results.	a) Maptask	b) Mapper	c) Task execution	d) All of the mentioned	A
4	function is responsible for consolidating the results produced by each of the Map() functions/tasks.	a) Reduce	b) Map	c) Reducer	d) All of the mentioned	A
5	Point out the wrong statement.	a) A MapReduce job usually splits	b) The MapReduce framework op	c) Applications typically impl	d) None of the mentioned	D
6	Although the Hadoop framework is implemented in Java, MapReduce applications need not be written in	a) Java	b) C	c) C#	d) None of the mentioned	A
7	is a utility which allows users to create and run jobs with any executables as the mapper and/or the reducer.	a) Hadoop Strdata	b) Hadoop Streaming	c) Hadoop Stream	d) None of the mentioned	B
8	maps input key/value pairs to a set of intermediate key/value pairs.	a) Mapper	b) Reducer	c) Both Mapper and Reducer	d) None of the mentioned	A
9	The number of maps is usually driven by the total size of	a) inputs	b) outputs	c) tasks	d) None of the mentioned	A
10	is the default Partitioner for partitioning key space.	a) HashPar	b) Partitioner	c) HashPartitioner	d) None of the mentioned	C

Q. No.	Question	Option A	Option B	Option C	Option D	Answer
1	Velocity of data refers to .	Speed of processing	Size of data	Complexity of data	None of the above	A
2	Variety of data refers to .	Different types and forms of data	Sources of data	Both A & B	None of the above	A
3	Rapid growth in data is referred as .	Data Conversion	Data diversion	Data explosion	All of the above	C
4	Which among the following is valid source of data?	Clickstream	Contracts	Application log	All of the above	D
5	Which of the following is not the phase of data analytics life cycle?	Communicate result	Model building	Model Planning	None of the above	D
6	The statement: big data can not be processed using traditional computing is .	TRUE	FALSE			A
7	Choose the correct statement: S1: Business intelligence requires structured data S2: Data science works with unstructured data.	S1 is true	S2 is true	Both S1 and S2 are true	Both S1 and S2 are false	C
8	Which of the following is not the application of big data?	Fraud detection	Call center analysis	IT log analytics	Traffic signal analysis	D
9	Which of the following is component of current analytical architecture?	Departmental warehouse	Dashboard	Data sources	All of the above	D
10	Which of the following is not the phase of data analytics life cycle?	Communicate result	Model building	Model Planning	None of the above	D
UNIT- I						
1	The rejection probability of Null Hypothesis when it is true is called as?	(a) Level of Confidence	(b) Level of Significance	(c) Level of Margin	(d) Level of Rejection	B
2	If the Critical region is evenly distributed then the test is referred as?	(a) Two tailed	(b) One tailed	(c) Three tailed	(d) Zero tailed	A
3	Consider a hypothesis H0 where $\phi_0 = 5$ against H1 where $\phi_1 > 5$. The test is?	(a) Right tailed	(b) Left tailed	(c) Center tailed	(d) Cross tailed	A
4	Type I error occurs when?	(a) We reject H0 if it is True	(b) We reject H0 if it is False	(c) We accept H0 if it is True	(d) We accept H0 if it is False	A
5	Analysis of variance is a statistical method of comparing the _____ of several populations.	(a). standard deviations	(b). variances	(c). means	(d). proportions	C
6	The _____ sum of squares measures the variability of the observed values around their respective treatment means.	(a). treatment	(b). error	(c). interaction	(d). total	B
7	If the true means of the k populations are equal, then MSTR/MSE should be:	(a). more than 1.00	(b). close to 1.00	(c). close to 0.00	(d). close to -1.00	B
8	To determine whether the test statistic of ANOVA is statistically significant, it can be compared to a critical value. What two pieces of information are needed to determine the critical value?	(a). sample size, number of groups	(b). mean, sample standard deviation	(c). expected frequency, obtained freq.	(d). MSTR, MSE	A
9	In ANOVA with 4 groups and a total sample size of 44, the computed F statistic is 2.33. In this case, the p-value is:	(a). exactly 0.05	(b). less than 0.05	(c). greater than 0.05	(d). cannot tell - it depends on what the SSE is	C
10	What is the function of a post-test in ANOVA?	(a). Determine if any statistically significant differences exist between the group means	(b). Describe those groups that have reliable differences	(c). Set the critical value for the F test	(d). None	B
UNIT- II						
1	The basic idea of the apriori algorithm is to generate item sets of a particular size & scans the database.	(A). candidate.	(B). primary.	(C). secondary.	(D). superkey.	A
2	is the most well known association rule algorithm and is used in most commercial products.	(A). Apriori algorithm.	(B). Partition algorithm.	(C). Distributed algorithm.	(D). Pincer-search algorithm.	A
3	An algorithm called _____ is used to generate the candidate item sets for each pass after the first.	(A). apriori.	(B). apriori-gen.	(C). sampling.	(D). partition.	B
4	The number of iterations in apriori	Increases with the size of data	Decreases with increase in size of data	increases with the size of max. frequent item	Decreases with increase in size of max. frequent item	C
5	In Apriori algorithm, if 1 item-sets are 100, then the number of candidate 2 item-sets are	4950	100	200	5000	A
6	Significant bottleneck in Apriori algorithm is	Finding frequent itemset	Pruning	Candidate Generation	Number of iterations	C
7	Logistic regression assumes a:	(A). Linear relationship between continuous variables	(B). Linear relationship between continuous variables	(C). Linear relationship between continuous variables	(D). Linear relationship between observations.	B
8	In binary logistic regression:	(A). The dependent variable is continuous.	(B). The dependent variable is divided into categories	(C). The dependent variable consists of binary values	(D). There is no dependent variable.	C
9	Which of the following is direct application of frequent itemset mining	Market basket analysis	Intrusion detection	Outlier detection	Social network analysis	A
UNIT- III						
1	Choose which data mining task is the most suitable for the following scenario: Given the records of books that a group of people read, find relationship of the genre pattern	Classification	Association rules	Clustering	Prediction	B
2	The basic idea of the apriori algorithm is to generate item sets of a particular size & scans the database.	(A). candidate.	(B). primary.	(C). secondary.	(D). superkey.	A
3	is the most well known association rule algorithm and is used in most commercial products.	(A). Apriori algorithm.	(B). Partition algorithm.	(C). Distributed algorithm.	(D). Pincer-search algorithm.	A
4	An algorithm called _____ is used to generate the candidate item sets for each pass after the first.	(A). apriori.	(B). apriori-gen.	(C). sampling.	(D). partition.	B
5	The number of iterations in apriori	Increases with the size of data	Decreases with increase in size of data	increases with the size of max. frequent item	Decreases with increase in size of max. frequent item	C
6	In Apriori algorithm, if 1 item-sets are 100, then the number of candidate 2 item-sets are	4950	100	200	5000	A
7	Significant bottleneck in Apriori algorithm is	Finding frequent itemset	Pruning	Candidate Generation	Number of iterations	C
8	Logistic regression assumes a:	(A). Linear relationship between continuous variables	(B). Linear relationship between continuous variables	(C). Linear relationship between continuous variables	(D). Linear relationship between observations.	B
9	In binary logistic regression:	(A). The dependent variable is continuous.	(B). The dependent variable is divided into categories	(C). The dependent variable consists of binary values	(D). There is no dependent variable.	C
10	Which of the following is direct application of frequent itemset mining	Market basket analysis	Intrusion detection	Outlier detection	Social network analysis	A
UNIT- IV						
1	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.	(a) Decision tree	(b) Graphs	(c) Trees	(d) Neural Networks	A
2	Decision Tree is a display of an algorithm.	(a) True	(b) False			A
3	Decision Trees can be used for Classification Tasks.	(a) True	(b) False			A
4	Choose from the following that are Decision Tree nodes?	(a) Decision Nodes	(b) End Nodes	(c) Chance Nodes	(d) All of the mentioned	D
5	Chance Nodes are represented by	(a) Disks	(b) Squares	(c) Circles	(d) Triangles	C
6	Given two Boolean random variables, A and B, where $P(A) = \frac{1}{2}$, $P(B) = 1/3$, and $P(A \neg B) = \frac{1}{4}$, what is $P(A B)$?	(a) 1/6	(b) 1/4	(c) 1/4	(d) 1	D
7	You've just finished training a decision tree for spam classification, and it's getting abnormally bad performance on both your training and test sets. You know that your implementation has no bugs, so what could be causing the problem?	(a) Your decision trees are too shallow.	(b) You need to increase the learning rate.	(c) You are overfitting.	(d) None of the above.	C
8	_____ refers to a model that can neither model the training data nor generalize to new data.	a) good fitting	b) overfitting	c) underfitting	d) all of the above	C
9	Which of the following is correct about the Naive Bayes?	(A) Assumes that all the features in a data set are independent.	(B) Assumes that all the features in a data set are dependent.	(C) Both	(D) All of the above	C
10	Naive Bayes Algorithm is a _____ learning algorithm.	(A) Supervised	(B) Reinforcement	(C) Unsupervised	(D) None of these	A
UNIT- V						
1	To visualize organization's formulation, communication etc. _____ type of visualization is used.	Statgeic	Metaphore	Concept	All of the above	A
2	Tabular representation takes _____ time to digest any information.	Lesser	Moderate	Longer	Very less	C
3	To summarize large datasets in the form of graphics _____ is used.	Tables	Histogram	Timeline	All of the above	B
4	Outliers can be easily identified using _____.	Bar charts	Pie charts	Line charts	Scatter Plots	D
5	Which of the following is type of data visualization?	Histogram	Timeline	Both A and B	None	C
6	The statement: Histograms are used to summarize tiny data sets in the form of tables is _____.	TRUE	FALSE			B
7	Scatter plots are also called as _____.	Histograms	Pie Charts	Heat maps	X- Y plots	C
8	To visualize organization's formulation, communication etc. _____ type of visualization is used.	Statgeic	Metaphore	Concept	All of the above	A
9	Tabular representation takes _____ time to digest any information.	Lesser	Moderate	Longer	Very less	C

10	Data visualization may be challenging due to .	Heterogeneity	Diversity	Both A & B	None of the above	C
1	is a platform for constructing data flows for extract, transform, and load (ETL) processing and analysis of large datasets.	a) Pig Latin b) Oozie c) Hive is not a relational database, but a	c) Pig d) Hive	d) All of the mentioned	C	
2	Point out the correct statement.	b) Hive is a relational database with SQL	c) Pig is a relational database with	d) All of the mentioned	A	
3	hides the limitations of Java behind a powerful and concise Clojure API for Cascading.	a) Cascalog	c) Cascalog	d) All of the mentioned	C	
4	Hive also support custom extensions written in	a) C# b) Java	c) C	d) C++	B	
5	Point out the wrong statement.	a) Elastic MapReduce (EMR) is Facebook	b) Amazon Web Service Elastic MapRed	c) Scalding is a Scala API on top o	d) All of the mentioned	A
6	is the most popular high-level Java API in Hadoop Ecosystem	a) Scalding b) HCatalog	c) Cascalog	d) Cascading	D	
7	is general-purpose computing model and runtime system for distributed data analytics.	a) Mapreduce b) Drill	c) Oozie	d) None of the mentioned	A	
8	The Pig Latin scripting language is not only a higher-level data flow language but also has operators similar to	a) SQL b) Mapreduce	c) XML	d) All of the mentioned	A	
9	jobs are optimized for scalability but not latency.	a) Mapreduce b) Drill	c) Oozie d) Hive	d) Chukwa	D	
10	is a framework for performing remote procedure calls and data serialization.	a) Drill b) BigTop	c) Avro	d) Chukwa	C	

NADAR SARASWATHI COLLEGE OF ENGINEERING AND TECHNOLOGY, THENI.

Course/Branch : B.E / CSE	Year / Semester : IVth YR / VII Sem	Format No.	NAC/TLP-07a.13
Subject Code : CS8091	Subject Name : Big Data Analytics	Rev. No.	02
Unit No : 01	Unit Name : Introduction to Big Data	Date	30.09.2020

OBJECTIVE TYPE QUESTION BANK

S. No.	Objective Questions (MCQ /True or False / Fill up with Choices)	BTL
1.	Which of the following is not an example of Social Media? a. Twitter b. Google c. Insta d. Youtube	L3
2.	By 2025, the volume of digital data will increase to a. TB b. YB c. ZB d. EB	L1
3.	For Drawing insights for Business what are need? a. Collecting the data b. Storing the data c. Analysing the data d. All the above	L5
4.	Does Facebook uses "Big Data " to perform the concept of Flashback? Is this True or False. a. TRUE b. FALSE	L3
5.	The Process of describing the data that is huge and complex to store and process is known as a. Analytics b. Data mining c. Big Data d. Data Warehouse	L1
6.	Data generated from online transactions is one of the example for volume of big data. Is this true or False. a. TRUE b. FALSE	L3
7.	Velocity is the speed at which the data is processed a. TRUE b. FALSE	L4
8.	_____ have a structure but cannot be stored in a database. a. Structured b. Semi-Structured c. Unstructured d. None of these	L2
9.	_____ refers to the ability to turn your data useful for business. a. Velocity b. Variety c. Value d. Volume	L1

NADAR SARASWATHI COLLEGE OF ENGINEERING AND TECHNOLOGY, THENI.

Course/Branch : B.E / CSE	Year / Semester : IVth YR / VII Sem	Format No.	NAC/TLP-07a.13
Subject Code : CS8091	Subject Name : Big Data Analytics	Rev. No.	02
Unit No : 01	Unit Name : Introduction to Big Data	Date	30.09.2020

OBJECTIVE TYPE QUESTION BANK

10.	Value tells the trustworthiness of data in terms of quality and accuracy. a. TRUE b. FALSE	L3
11.	GFS consists of a _____ Master and _____ Chunk Servers a. Single, Single b. Multiple, Single c. Single, Multiple d. Multiple, Multiple	L1
12.	Files are divided into _____ sized Chunks. a. Static b. Dynamic c. Fixed d. Variable	L2
13.	_____ is an open source framework for storing data and running application on clusters of commodity hardware. a. HDFS b. Hadoop c. MapReduce d. Cloud	L1
14.	HDFS Stores how much data in each clusters that can be scaled at any time? a. 32 b. 64 c. 128 d. 256	L2
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 a. TRUE b. FALSE	L4
16.	Hortonworks was introduced by Cloudera and owned by Yahoo. a. TRUE b. FALSE	L1
17.	Hadoop YARN is used for Cluster Resource Management in Hadoop Ecosystem. a. TRUE b. FALSE	L4
18.	Google Introduced MapReduce Programming model in 2004. a. TRUE b. FALSE	L4
19.	_____ phase sorts the data & _____ creates logical clusters. a. Reduce, YARN b. MAP, YARN c. REDUCE, MAP d. MAP, REDUCE	L2

NADAR SARASWATHI COLLEGE OF ENGINEERING AND TECHNOLOGY, THENI.

Course/Branch : B.E / CSE	Year / Semester : IVth YR / VII Sem	Format No.	NAC/TLP-07a.13
Subject Code : CS8091	Subject Name : Big Data Analytics	Rev. No.	02
Unit No : 01	Unit Name : Introduction to Big Data	Date	30.09.2020

OBJECTIVE TYPE QUESTION BANK

20.	There is only one operation between Mapping and Reducing is it True or False... a. TRUE b. FALSE	L4
21.	_____ is factors considered before Adopting Big Data Technology. a. Validation b. Verification c. Data d. Design	L3
22.	_____ for improving supply chain management to optimize stock management, replenishment, and forecasting; a. Descriptive b. Diagnostic c. Predictive d. Prescriptive	L3
23.	which among the following is not a Data mining and analytical applications? a. profile matching b. social network analysis c. facial recognition d. Filtering	L2
24.	_____ as a result of data accessibility, data latency, data availability, or limits on bandwidth in relation to the size of inputs. a. Computation-restricted throttling b. Large data volumes c. Data throttling d. Benefits from data parallelization	L1
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. a. Lowering costs b. Increasing revenues c. Increasing productivity d. Reducing risk	L2
26.	Which storage subsystem can support massive data volumes of increasing size. a. Extensibility b. Fault tolerance c. Scalability d. High-speed I/O capacity	L5
27.	_____ provides performance through distribution of data and fault tolerance through replication a. HDFS b. PIG c. HIVE d. HADOOP	L3

NADAR SARASWATHI COLLEGE OF ENGINEERING AND TECHNOLOGY, THENI.

Course/Branch : B.E / CSE	Year / Semester : IVth YR / VII Sem	Format No.	NAC/TLP-07a.13
Subject Code : CS8091	Subject Name : Big Data Analytics	Rev. No.	02
Unit No : 01	Unit Name : Introduction to Big Data	Date	30.09.2020

OBJECTIVE TYPE QUESTION BANK

28.	<p>_____ is a programming model for writing applications that can process Big Data in parallel on multiple nodes.</p> <ul style="list-style-type: none"> a. HDFS b. MAP REDUCE c. HADOOP d. HIVE 	L1
29.	<p>_____ takes the grouped key-value paired data as input and runs a Reducer function on each one of them.</p> <ul style="list-style-type: none"> a. MAPPER b. REDUCER c. COMBINER d. PARTITIONER 	L2
30.	<p>_____ is a type of local Reducer that groups similar data from the map phase into identifiable sets.</p> <ul style="list-style-type: none"> a. MAPPER b. REDUCER c. COMBINER d. PARTITIONER 	L3
31.	<p>While Installing Hadoop how many xml files are edited and list them ?</p> <ul style="list-style-type: none"> i. core-site.xml ii. hdfs-site.xml iii. mapred.xml iv. yarn.xml 	L4
32.	<p>Write the code for core-site.xml ?</p> <pre><?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 ></pre>	L6
33.	Write the code for hdfs-site.xml ?	L3

NADAR SARASWATHI COLLEGE OF ENGINEERING AND TECHNOLOGY, THENI.

Course/Branch : B.E / CSE	Year / Semester : IVth YR / VII Sem	Format No.	NAC/TLP-07a.13
Subject Code : CS8091	Subject Name : Big Data Analytics	Rev. No.	02
Unit No : 01	Unit Name : Introduction to Big Data	Date	30.09.2020

OBJECTIVE TYPE QUESTION BANK

	<pre><?xml version="1.0" encoding="UTF-8"?> <?xml-stylesheet type="text/xsl" href="configuration.xsl"?> <configuration> <property><name>dfs.replication</name><value>1</value></property> <property> <name>dfs.namenode.name.dir</name><value>/hadoop2.6.0/data/name</value><final>true</final></property> <property><name>dfs.datanode.data.dir</name><value>/hadoop2.6.0/data/data</value><final>true</final> </property> </configuration> </xml></pre>	
34.	<p>Write the code for mapred.xml?</p> <pre><?xml version="1.0"?> <configuration> <property> <name>mapreduce.framework.name</name> <value>yarn</value> </property> <property> <name>mapred.job.tracker</name> <value>localhost:9001</value> </property> <property> <name>mapreduce.application.classpath</name> <value>/hadoop-2.6.0/share/hadoop/mapreduce/*, /hadoop-2.6.0/share/hadoop/mapreduce/lib/*, /hadoop-2.6.0/share/hadoop/common/*, /hadoop-2.6.0/share/hadoop/common/lib/*, /hadoop-2.6.0/share/hadoop/yarn/*, /hadoop-2.6.0/share/hadoop/yarn/lib/*, /hadoop-2.6.0/share/hadoop/hdfs/*, /hadoop-2.6.0/share/hadoop/hdfs/lib/*, </value> </property> </configuration></pre>	L3
35.	<p>Write the code for yarn-site.xml ?</p> <pre><?xml version="1.0"?> <configuration> <property> <name>yarn.nodemanager.aux-services</name> <value>mapreduce_shuffle</value> </property></pre>	L3

NADAR SARASWATHI COLLEGE OF ENGINEERING AND TECHNOLOGY, THENI.

Course/Branch : B.E / CSE	Year / Semester : IVth YR / VII Sem	Format No.	NAC/TLP-07a.13
Subject Code : CS8091	Subject Name : Big Data Analytics	Rev. No.	02
Unit No : 01	Unit Name : Introduction to Big Data	Date	30.09.2020

OBJECTIVE TYPE QUESTION BANK

	<pre> <property> <name>yarn.nodemanager.aux-services.mapreduce.shuffle.class</name> <value>org.apache.hadoop.mapred.ShuffleHandler</value> </property> <property> <name>yarn.nodemanager.log-dirs</name> <value>D:\hadoop\userlog</value><final>true</final> </property> <property><name>yarn.nodemanager.local- dirs</name><value>D:\hadoop\temp\nm-localdir</value></property> <property> <name>yarn.nodemanager.delete.debug-delay-sec</name> <value>600</value> </property> <property><name>yarn.application.classpath</name> <value>/hadoop-2.6.0/,/hadoop- 2.6.0/share/hadoop/common/*,/hadoop2.6.0/share/hadoop/common/lib/*,/hadoop- 2.6.0/share/hadoop/hdfs/*,/hadoop2.6.0/share/hadoop/hdfs/lib/*,/hadoop- 2.6.0/share/hadoop/mapreduce/*,/hadoop2.6.0/share/hadoop/mapreduce/lib/*,/hado- op-2.6.0/share/hadoop/yarn/*,/hadoop2.6.0/share/hadoop/yarn/lib/*</value> </property> </configuration> </pre>	
36.	<p>what are the environmental variable set for Hadoop ?</p> <p>i. User variables:</p> <ul style="list-style-type: none"> • Variable: HADOOP_HOME • Value: D:\hadoop-2.6.0 <p>ii. System variable:</p> <ul style="list-style-type: none"> • Variable: Path • Value: D:\hadoop-2.6.0\bin D:\hadoop-2.6.0\sbin D:\hadoop-2.6.0\share\hadoop\common* D:\hadoop-2.6.0\share\hadoop\hdfs D:\hadoop-2.6.0\share\hadoop\hdfs\lib* D:\hadoop-2.6.0\share\hadoop\hdfs* D:\hadoop-2.6.0\share\hadoop\yarn\lib* D:\hadoop-2.6.0\share\hadoop\yarn* D:\hadoop-2.6.0\share\hadoop\mapreduce\lib* D:\hadoop-2.6.0\share\hadoop\mapreduce* D:\hadoop-2.6.0\share\hadoop\common\lib* 	L1

NADAR SARASWATHI COLLEGE OF ENGINEERING AND TECHNOLOGY, THENI.

Course/Branch : B.E / CSE	Year / Semester : IVth YR / VII Sem	Format No.	NAC/TLP-07a.13
Subject Code : CS8091	Subject Name : Big Data Analytics	Rev. No.	02
Unit No : 02	Unit Name : Clustering and Classification	Date	30.09.2020

OBJECTIVE TYPE QUESTION BANK

S. No.	Objective Questions (MCQ /True or False / Fill up with Choices)	BTL
1.	Movie Recommendation systems are an example of 1. Classification 2. Clustering 3. Reinforcement Learning 4. Regression a. 2 Only b. 1 and 2 c. 1 and 3 d. 2 and 3	L3
2.	Sentiment Analysis is an example of 1. Regression 2. Classification 3. Clustering 4 Reinforcement Learning a. 1, 2 and 4 b. 1 and 3 c. 1, 2 and 3 d. 1 and 2	L3
3.	Can decision trees be used for performing clustering? a. True b. False	L4
4.	What is the minimum no. of variables/ features required to perform clustering? 1. 0 2. 1 3. 2 4. 3	L1
5.	For two runs of K-Mean clustering is it expected to get same clustering results? 1. Yes 2. No	L3
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. a. 1, 3 and 4 b. 1, 2 and 3 c. 1, 2 and 4 d. All of the above	L1
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	L3
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?	L6

NADAR SARASWATHI COLLEGE OF ENGINEERING AND TECHNOLOGY, THENI.

Course/Branch : B.E / CSE	Year / Semester : IVth YR / VII Sem	Format No.	NAC/TLP-07a.13
Subject Code : CS8091	Subject Name : Big Data Analytics	Rev. No.	02
Unit No : 02	Unit Name : Clustering and Classification	Date	30.09.2020

OBJECTIVE TYPE QUESTION BANK

	<p>a. There were 28 data points in clustering analysis b. The best no. of clusters for the analyzed data points is 4 c. The proximity function used is Average-link clustering d. The above dendrogram interpretation is not possible for K-Means clustering analysis</p>	
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?	L6
	<p>1. 1 2. 2 3. 3 4. 4</p>	
10.	<p>In which of the following cases will K-Means clustering fail to give good results?</p> <ol style="list-style-type: none"> Data points with outliers Data points with different densities Data points with round shapes Data points with non-convex shapes <p>a. 1 and 2 b. 2 and 3 c. 2 and 4 d. 1, 2 and 4</p>	L4
11.	<p>The discrete variables and continuous variables are two types of</p> <ol style="list-style-type: none"> Open end classification Time series classification Qualitative classification Quantitative classification 	L1

NADAR SARASWATHI COLLEGE OF ENGINEERING AND TECHNOLOGY, THENI.

Course/Branch : B.E / CSE	Year / Semester : IVth YR / VII Sem	Format No.	NAC/TLP-07a.13
Subject Code : CS8091	Subject Name : Big Data Analytics	Rev. No.	02
Unit No : 02	Unit Name : Clustering and Classification	Date	30.09.2020

OBJECTIVE TYPE QUESTION BANK

12.	<p>Bayesian classifiers is</p> <ol style="list-style-type: none"> 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 	L1
13.	<p>Classification accuracy is</p> <ol style="list-style-type: none"> 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 	L1
14.	<p>Classification task referred to</p> <ol style="list-style-type: none"> 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 	L1
15.	<p>Euclidean distance measure is</p> <ol style="list-style-type: none"> 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 	L1
16.	<p>_____ is good at handle missing data and support both the kind of attributes (i.e Categorical and Continuous attributes)</p> <ol style="list-style-type: none"> a. ID3. b. C4.5. c. CART. d. Naïve Bayes. 	L4
17.	<p>Decision trees use _____, in that they always choose the option that seems the best available at that moment.</p> <ol style="list-style-type: none"> a. Greedy Algorithms. b. Divide and Conquer. c. Backtracking. d. Shortest Path Method. 	L2
18.	<p>Decision trees cannot handle categorical attributes with many distinct values, such as country codes for telephone numbers.</p> <ol style="list-style-type: none"> a. TRUE b. FALSE 	L4
19.	<p>_____ are easy to implement and can execute efficiently even without</p>	L2

NADAR SARASWATHI COLLEGE OF ENGINEERING AND TECHNOLOGY, THENI.

Course/Branch : B.E / CSE	Year / Semester : IVth YR / VII Sem	Format No.	NAC/TLP-07a.13
Subject Code : CS8091	Subject Name : Big Data Analytics	Rev. No.	02
Unit No : 02	Unit Name : Clustering and Classification	Date	30.09.2020

OBJECTIVE TYPE QUESTION BANK

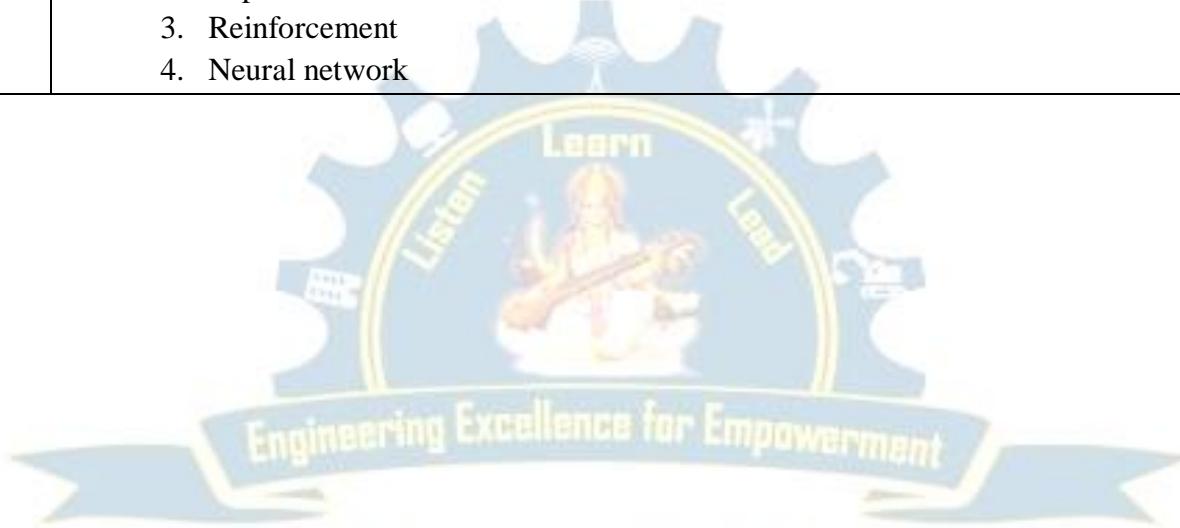
	prior knowledge of the data, they are among the most popular algorithms for classifying text documents. a. ID3 b. Naïve Bayes classifiers c. CART d. None of these.	
20.	High entropy means that the partitions in classification are a. Pure b. Not pure c. Useful d. Useless	L2
21.	Which of the following statements about Naive Bayes is incorrect? a. Attributes are equally important. b. Attributes are statistically dependent of one another given the class value. c. Attributes are statistically independent of one another given the class value. d. Attributes can be nominal or numeric	L4
22.	The maximum value for entropy depends on the number of classes so if we have 8 Classes what will be the max entropy. $\text{Entropy}(x) = - \sum (P(x=k) * \log_2(P(x=k)))$ a. Max Entropy is 1 b. Max Entropy is 2 c. Max Entropy is 3 d. Max Entropy is 4	L6
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	L6
24.	Point out the wrong statement. a. k-nearest neighbor is same as k-means b. k-means clustering is a method of vector quantization c. k-means clustering aims to partition n observations into k clusters d. none of the mentioned	L4
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	L3

NADAR SARASWATHI COLLEGE OF ENGINEERING AND TECHNOLOGY, THENI.

Course/Branch : B.E / CSE	Year / Semester : IVth YR / VII Sem	Format No.	NAC/TLP-07a.13
Subject Code : CS8091	Subject Name : Big Data Analytics	Rev. No.	02
Unit No : 02	Unit Name : Clustering and Classification	Date	30.09.2020

OBJECTIVE TYPE QUESTION BANK

	3. Regression 4. None of These	
26.	Can we use K Mean Clustering to identify the objects in video? 1. Yes 2. No	L4
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	L2



Nadar Saraswathi College of
Engineering & Technology

NADAR SARASWATHI COLLEGE OF ENGINEERING AND TECHNOLOGY, THENI.

Course/Branch : B.E / CSE	Year / Semester : IVth YR / VII Sem	Format No.	NAC/TLP-07a.13
Subject Code : CS8091	Subject Name : Big Data Analytics	Rev. No.	02
Unit No : 03	Unit Name : Association and Recommendation	Date	30.09.2020

OBJECTIVE TYPE QUESTION BANK

S. No.	Objective Questions (MCQ /True or False / Fill up with Choices)	BTL
1.	<p>_____ metric is examined to determine a reasonably optimal value of k.</p> <ol style="list-style-type: none"> 1. Mean Square Error 2. Within Sum of Squares (WSS) 3. Speed 4. None of These 	L5
2.	<p>If an itemset is considered frequent, then any subset of the frequent itemset must also be frequent.</p> <ol style="list-style-type: none"> 1. Apriori Property 2. Downward Closure Property 3. Either 1 or 2 4. Both 1 & 2 	L2
3.	<p>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</p> <ol style="list-style-type: none"> 1. 0 2. 1 3. 2 4. 3 	L6
4.	<p>Confidence is a measure of how X and Y are really related rather than coincidentally happening together.</p> <ol style="list-style-type: none"> a. True b. False 	L4
5.	<p>A high-confidence rule can sometimes be misleading because confidence does not consider support of the itemset in the rule consequent. Is This True ?</p> <ol style="list-style-type: none"> a. Yes b. No 	L4
6.	<p>_____ recommend items based on similarity measures between users and/or items.</p> <ol style="list-style-type: none"> 1. Content Based Systems 2. Hybrid System 3. Collaborative Filtering Systems 4. None of These 	L2
7.	<p>There are _____ major Classification of Collaborative Filtering Mechanisms</p> <ol style="list-style-type: none"> 1. 1 2. 2 3. 3 4. None of These 	L1
8.	<p>Movie Recommendation to peoples is an example of</p> <ol style="list-style-type: none"> 1. User Based Recommendation 2. Item Based Recommendation 3. Knowledge Based Recommendation 4. Content Based Recommendation 	L3

NADAR SARASWATHI COLLEGE OF ENGINEERING AND TECHNOLOGY, THENI.

Course/Branch : B.E / CSE	Year / Semester : IVth YR / VII Sem	Format No.	NAC/TLP-07a.13
Subject Code : CS8091	Subject Name : Big Data Analytics	Rev. No.	02
Unit No : 03	Unit Name : Association and Recommendation	Date	30.09.2020

OBJECTIVE TYPE QUESTION BANK

9.	<p>_____ recommenders rely on an explicitly defined set of recommendation rules.</p> <ol style="list-style-type: none"> 1. Constraint Based 2. Case Based 3. Content Based 4. User Based 	L2
10.	<p>Parallelized hybrid recommender systems operate independently of one another and produce separate recommendation lists.</p> <ol style="list-style-type: none"> 1. True 2. False 	L4
11.	<p>Association rules are sometimes referred to as</p> <ol style="list-style-type: none"> a. market basket analysis b. Itemset Filtering c. Frequent Itemset Analysis d. None of these. 	L1
12.	<p>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</p> <ol style="list-style-type: none"> a. 0.4 b. 0.5 c. 0.6 d. 0.7 	L6
13.	<p>Lift is defined as the measure of certainty or trustworthiness associated with each discovered rule.</p> <ol style="list-style-type: none"> a. TRUE b. FALSE 	L4
14.	<p>_____ is able to identify trustworthy rules, but it cannot tell whether a rule is coincidental.</p> <ol style="list-style-type: none"> a. Lift b. Confidence c. Support d. Leverage 	L1
15.	<p>_____ recommend items based on similarity measures between users and/or items. The items recommended to a user are those preferred by similar users.</p> <ol style="list-style-type: none"> a. Collaborative Filtering System b. Content Based Recommendation c. Knowledge Based Recommendation d. Hybrid Approaches 	L2
16.	<p>Pure collaborative approaches take a matrix of given user-item ratings as the only input and typically produce output. Is it Pure Collaborative?</p> <ol style="list-style-type: none"> a. Yes b. No 	L4
17.	<p>With respect to the determination of the set of similar users, one common measure used in recommender systems is</p>	L1

NADAR SARASWATHI COLLEGE OF ENGINEERING AND TECHNOLOGY, THENI.

Course/Branch : B.E / CSE	Year / Semester : IVth YR / VII Sem	Format No.	NAC/TLP-07a.13
Subject Code : CS8091	Subject Name : Big Data Analytics	Rev. No.	02
Unit No : 03	Unit Name : Association and Recommendation	Date	30.09.2020

OBJECTIVE TYPE QUESTION BANK

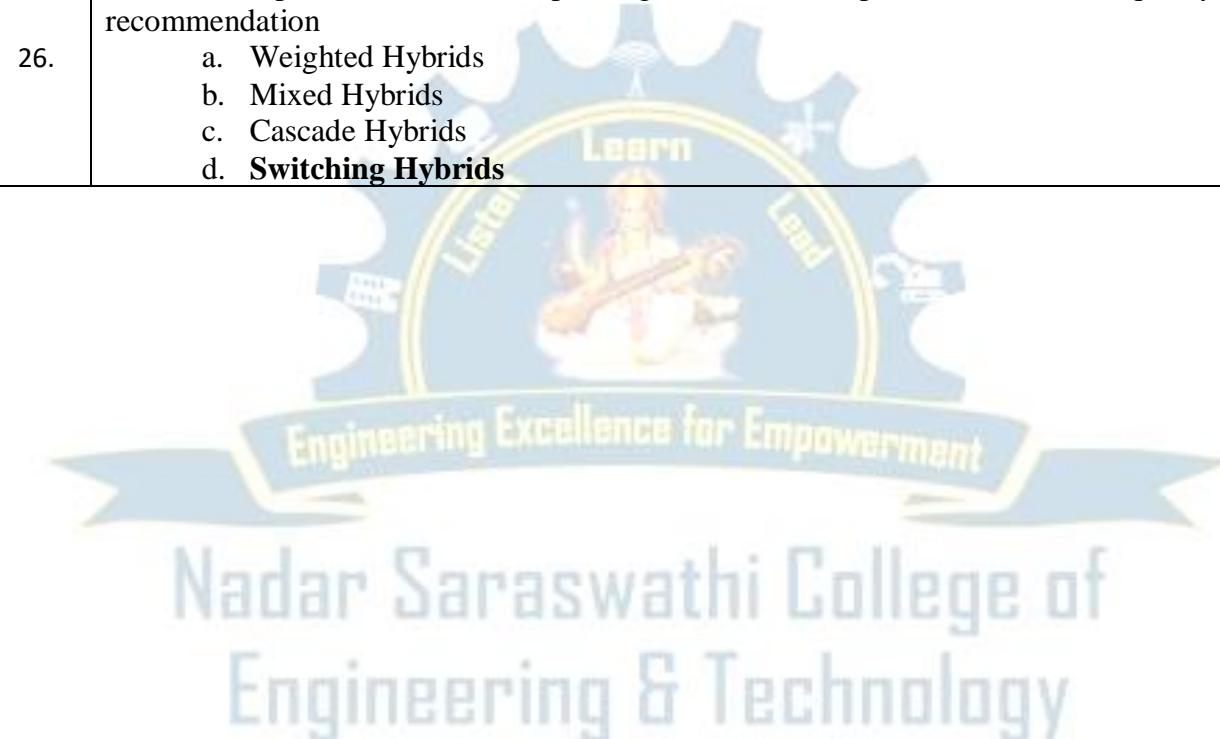
	a. Cosine Similarity Measure b. Pearson's correlation coefficient. c. Mean Squared Error Method d. None of these.	
18.	Large-scale e-commerce sites, often implement a different technique, _____ which is more apt for offline preprocessing and thus allows for the computation of recommendations in real time even for a very large rating matrix. a. Item-Based Recommendation b. User-Based Recommendation c. Content-Based Recommendation d. None of these	L2
19.	Here are two very short texts to compare and find the cosine similarity measure? I. Julie loves me more than Linda loves me II. Jane likes me more than Julie loves me a. 0.6 b. 0.7 c. 0.8 d. 0.9	L6
20.	_____ is based on the availability of item descriptions and a profile that assigns importance to these characteristics. a. Item-Based Recommendation b. User-Based Recommendation c. Content-Based Recommendation. d. None of these	L2
21.	Consider the features of a movie which are not relevant to a recommendation system. a. The set of actors of the movie. b. The Director c. The Year in which the movie was made d. The Budget of the movie.	L3
22.	A _____ has been implemented, for similarity based retrieval under nearest neighbors. a. k-nearest-neighbor method (kNN) b. Conventional Neural Network (CNN) c. Bayes Theorem d. Naïve Bayes Classifier	L2
23.	Case-based recommenders focus on the retrieval of similar items on the basis of different types of similarity measures a. TRUE b. FALSE	L4
24.	In _____ recommendation approaches, items are retrieved using similarity measures that describe to which extent item properties match some given user's requirements. a. Item-Based b. Case-Based	L2

NADAR SARASWATHI COLLEGE OF ENGINEERING AND TECHNOLOGY, THENI.

Course/Branch : B.E / CSE	Year / Semester : IVth YR / VII Sem	Format No.	NAC/TLP-07a.13
Subject Code : CS8091	Subject Name : Big Data Analytics	Rev. No.	02
Unit No : 03	Unit Name : Association and Recommendation	Date	30.09.2020

OBJECTIVE TYPE QUESTION BANK

	c. Content-Based d. User-Based	
25.	_____ are based on a sequenced order of techniques, in which each succeeding recommender only refines the recommendations of its predecessor. a. Weighted Hybrids b. Mixed Hybrids c. Cascade Hybrids d. Switching Hybrids	L1
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 a. Weighted Hybrids b. Mixed Hybrids c. Cascade Hybrids d. Switching Hybrids	L1



NADAR SARASWATHI COLLEGE OF ENGINEERING AND TECHNOLOGY, THENI.

Course/Branch : B.E / CSE	Year / Semester : IVth YR / VII Sem	Format No.	NAC/TLP-07a.13
Subject Code : CS8091	Subject Name : Big Data Analytics	Rev. No.	02
Unit No : 04	Unit Name : Stream Concepts	Date	30.09.2020

OBJECTIVE TYPE QUESTION BANK

S. No.	Objective Questions (MCQ /True or False / Fill up with Choices)	BTL
1.	Which one doesn't belong to application of Data Stream? 1. Web traffic 2. Internet 3. Sensor data 4. None of these	L1
2.	These queries are, in a sense, permanently executing and produce outputs at appropriate times. a. Standing query b. Ad-hoc query	L1
3.	Google wants to know what queries are more frequent today than yesterday. 1. Mining Query Stream 2. Mining Login Stream 3. Mining Search Stream 4. Mining Click Stream	L4
4.	Yahoo wants to know which of its pages are getting an unusual number of hits in the past hour 1. Mining Query Stream 2. Mining Login Stream 3. Mining Search Stream 4. Mining Click Stream	L4
5.	Surveillance cameras produce images with high resolution than satellites 1. True 2. False	L4
6.	In the data stream model, individual data items may be _____, e.g., network measurements, call records, web page visits, sensor readings, and so on. a. Key – Value Pair b. Relational Tuples c. Variable d. Database	L2
7.	A data stream is a real time continuous and ordered sequence of items. It is possible to control the order in which the items arrive, nor it is feasible to locally store a stream in its entirety in any memory device. Is this Statement True... a. YES b. NO	L4
8.	Long running queries are registered in the _____ and placed into groups for shared processing. a. Query Repository b. Archival Storage c. Limited Working Storage	L1

NADAR SARASWATHI COLLEGE OF ENGINEERING AND TECHNOLOGY, THENI.

Course/Branch : B.E / CSE	Year / Semester : IVth YR / VII Sem	Format No.	NAC/TLP-07a.13
Subject Code : CS8091	Subject Name : Big Data Analytics	Rev. No.	02
Unit No : 04	Unit Name : Stream Concepts	Date	30.09.2020

OBJECTIVE TYPE QUESTION BANK

	d. Main Memory	
9.	Which among the following was not the example of Data Stream concepts? a. Financial Applications b. Network Monitoring c. Fraud Detection d. Web Application	L3
10.	In DSMS, The data model and query processor will allow either order based or time based operations. Is this Statement True.. a. YES b. NO	L4
11.	In Streaming Queries , Alerting the user when Stock crosses over a price point is an example of _____ a. Continuous Queries b. One Time Queries c. Sampling Queries d. None of These	L1
12.	For Example : An increase in queries like “dengue fever symptoms” enables us to predict the number of sufferers. Which one it belong. a. Query Stream b. Click Stream c. User Stream d. Content Stream	L3
13.	Standing Queries are executed when user prefer and produce output at appropriate times... Is This True ? a. YES b. NO	L4
14.	At Ocean Surface Temperature Sensor, the data stream model will process the maximum temperature ever recorded analysis is this? a. Standing Query b. Ad-hoc Query	L2
15.	A useful model of stream processing is that queries are about a window of length N – the N most recent elements received. N is so large it cannot be stored in memory, or even on disk. Is this Statement is ? a. TRUE b. FALSE	L4
16.	The stream-processing algorithm is executed in query processor, without access to main memory or with only rare accesses to secondary storage. Is this Statement True? a. YES b. NO	L3
17.	Web sites often like to report the number of unique users over the past month. Kindly complete the sql : SELECT _____(_____(name)) FROM Logins	L6

NADAR SARASWATHI COLLEGE OF ENGINEERING AND TECHNOLOGY, THENI.

Course/Branch : B.E / CSE	Year / Semester : IVth YR / VII Sem	Format No.	NAC/TLP-07a.13
Subject Code : CS8091	Subject Name : Big Data Analytics	Rev. No.	02
Unit No : 04	Unit Name : Stream Concepts	Date	30.09.2020

OBJECTIVE TYPE QUESTION BANK

	WHERE time_____ t; a. SUM, UNIQUE, == b. COUNT,UNIQUE,<= c. COUNT, DISTINCT, >= d. SUM, DISTINCT,<=	
18.	A search engine receives a stream of queries, and it would like to study the behavior of typical users. Assume that the stream consists of tuples (user, query, time) a. Standing Query b. Ad-hoc Query	L3
19.	More generally, we can obtain a sample consisting of any rational fraction a/b of the users by hashing user names to b buckets, 0 through $b - 1$. Add the search query to the sample if the hash value is less than a . Is this Statement True? a. YES b. NO	L4
20.	Which filtering eliminate most of the tuples that do not meet the criteria? a. Blooms Filtering b. AMS Filtering c. DGIM Filtering d. None of These	L5
21.	_____ is reactive because it waits for users to request a query and then delivers the analytics. a. On Demand Real Time Analytics b. Continuous Real Time Analytics c. Time based Analytics d. Content Based Analytics	L3
22.	Monitoring stock market trends provide analytics to help users make a decision to buy or sell all in real time. a. On Demand Real Time Analytics b. Continuous Real Time Analytics c. Time based Analytics d. Content Based Analytics	L3
23.	Sentiment analysis is widely applied to reviews and social media for a variety of applications ranging from marketing to customer service. Is it True... a. TRUE b. FALSE	L2
24.	_____ is based on a model of representing individual entities and numerous kinds of relationships that connect those entities. a. Graph analytics b. Real time analytics c. Sentiment Analysis d. Stock market Prediction	L1
25.	A _____ that can be represented using a triple format consisting of a subject (the source point of the relationship), an object (the target), and a predicate	L2

NADAR SARASWATHI COLLEGE OF ENGINEERING AND TECHNOLOGY, THENI.

Course/Branch : B.E / CSE	Year / Semester : IVth YR / VII Sem	Format No.	NAC/TLP-07a.13
Subject Code : CS8091	Subject Name : Big Data Analytics	Rev. No.	02
Unit No : 04	Unit Name : Stream Concepts	Date	30.09.2020

OBJECTIVE TYPE QUESTION BANK

(that models the type of the relationship).	
a. Directed Graph b. Undirected Graph c. Weighted Graph d. Un Weighted Graph	



NADAR SARASWATHI COLLEGE OF ENGINEERING AND TECHNOLOGY, THENI.

Course/Branch : B.E / CSE	Year / Semester : IVth YR / VII Sem	Format No.	NAC/TLP-07a.13
Subject Code : CS8091	Subject Name : Big Data Analytics	Rev. No.	02
Unit No : 05	Unit Name : NoSQL Data Management for Big Data and Visualization	Date	30.09.2020

OBJECTIVE TYPE QUESTION BANK

S. No.	Objective Questions (MCQ /True or False / Fill up with Choices)	BTL
1.	Among top 10 ranking database model with relational DBMS which one is First ? a. MySQL b. Oracle c. MongoDB d. Cassandra	L1
2.	Which among the Database which one is popular ? a. MongoDB b. Oracle c. MySQL d. Cassandra	L2
3.	Which among the following are incorrect in regards with NoSQL ? a. Its Easy and ready to manage with clusters. b. Suitable for upcoming data explosions. c. It requires to keep track with data structure d. Provide easy and flexible system.	L4
4.	Which Database Administrator job was in trends with job trends ? a. MongoDB b. CouchDB c. SimpleDB d. Redis	L2
5.	No SQL Means _____ a. Not SQL b. No Usage of SQL c. Not Only SQL d. Not for SQL	L1
6.	In Relational database Management System Scaling is possible a. TRUE b. FALSE	L4
7.	Which among the following is not the example of NoSql ? a. Google b. NetFlix c. Amazon d. CERN	L3
8.	Carlo Strozzi used the term NoSQL in _____ to name his lightweight, open-source relational database that did not expose the standard SQL interface. a. 1965 b. 1989 c. 1998 d. 2007	L1
9.	In Brewer's Cap Theorem which among the following was not considered ? a. Consistency	L2

NADAR SARASWATHI COLLEGE OF ENGINEERING AND TECHNOLOGY, THENI.

Course/Branch : B.E / CSE	Year / Semester : IVth YR / VII Sem	Format No.	NAC/TLP-07a.13
Subject Code : CS8091	Subject Name : Big Data Analytics	Rev. No.	02
Unit No : 05	Unit Name : NoSQL Data Management for Big Data and Visualization	Date	30.09.2020

OBJECTIVE TYPE QUESTION BANK

	b. Availability c. Partition Tolerance d. None of these	
10.	“If the network is broken, your database won’t work “because RDBMS have Network partitions. Is this statement True? a. Yes b. No	L4
11.	If CAP Considers only Availability and Partition what are apt example in real-time ? a. BigTable b. Dynamo c. Postgres d. None of these	L1
12.	If CAP Considers only Consistency and Partition what are apt example in real-time ? a. BigTable b. Dynamo c. Postgres d. None of these	L1
13.	If CAP Considers only Availability and Availability what are apt example in real-time ? a. BigTable b. Dynamo c. Postgres d. None of these	L1
14.	Scalability and better performance of No SQL is Achieved by sacrificing ACID Compatibility Is it TRUE? a. TRUE b. FALSE	L4
15.	In Document Based NoSQL, All Documents are usually organized into collections or databases with unique structure. Is this True ? a. TRUE b. FALSE	L4
16.	Key Value store was used in which real time applications ? a. BigTable b. Dynamo c. Postgres d. None of these	L1
17.	Graph model of NOSQL was used in ? a. Twitter b. Facebook c. Google d. Whatsapp	L3
18.	Column Based Model of NoSQL was not supported in ? a. Twitter b. Facebook	L3

NADAR SARASWATHI COLLEGE OF ENGINEERING AND TECHNOLOGY, THENI.

Course/Branch : B.E / CSE	Year / Semester : IVth YR / VII Sem	Format No.	NAC/TLP-07a.13
Subject Code : CS8091	Subject Name : Big Data Analytics	Rev. No.	02
Unit No : 05	Unit Name : NoSQL Data Management for Big Data and Visualization	Date	30.09.2020

OBJECTIVE TYPE QUESTION BANK

	c. Google d. BigTable	
19.	Document Based Model was used in ? a. MongoDB b. CouchDB c. SimpleDB d. Redis	L3
20.	MongoDB is _____ a. Column Based b. Key Value Based c. Document Based d. Graph Based	L2
21.	_____ is the process of storing data records across multiple machines a. Sharding b. HDFS c. HIVE d. HBASE	L1
22.	The results of a hive query can be stored as a. Local File b. HDFS File c. Both d. Cannot be stored	L2
23.	The position of a specific column in a Hive table a. can be anywhere in the table creation clause b. must match the position of the corresponding data in the data file c. Must match the position only for date time data type in the data file d. Must be arranged alphabetically	L3
24.	The Hbase tables are A. Made read only by setting the read-only option B. Always writeable C. Always read-only D. Are made read only using the query to the table	L1
25.	Hbase creates a new version of a record during A. Creation of a record B. Modification of a record C. Deletion of a record D. All the above	L2