Business Data Analytics Methods and Tools

1. Identify which of the following is a drawback of filling in a global constant for the missing value?

   A. It may project wrong trend in data

   B. It decreases the number of missing values

   C. It increases the data size

   D. It is difficult to update the data

ANSWER: A

2. Choose the wrong data mining functionality among the given data mining functionalities.

   A. Classification

   B. Object Description

   C. Class Description

   D. Clustering

ANSWER: B

3. Identify the types of data, which cannot be used for mining?

   A. File System data

   B. Data warehouse data

   C. Database data

   D. Transactional data

ANSWER: A

4. Identify among the following methods which is not used for constructing a neural network?

   A. Stemming Algorithms

   B. Support Vector Machines

   C. Bayesian Classification

   D. K-nearest neighbor Classification

ANSWER: A

5. Choose from the following, which is not true about mode?

   A. A data set can have at most one mode

   B. It is defined for both qualitative and quantitative values

   C. It is the most frequently occurring value

   D. A data set can have more than one mode

ANSWER: A

6. What it is called when data points that partition a data distribution into consecutive groups of equal size?

   A. Intervals

   B. Whiskers

   C. Quantiles

   D. Range

ANSWER: C

7. The median is also called as _____

   A. 5-quantile

   B. 3-quantile

   C. 4-quantile

   D. 2-quantile

ANSWER: D

8. The 4-quantiles are also called as?

   A. Percentiles

   B. Octiles

   C. Quintiles

   D. Quartiles

ANSWER: D

9. What is true about ETL (Extraction, Transformation, Loading) tools?

   A. Are not used for data transformations

   B. Do not allow the use of graphical user interface

   C. Do not allow data transformations

   D. Allow the use of graphical user interface

ANSWER: D

10. Identify which of the following is contained in NumPy library?

    A. n-dimensional array object

    B. all of the mentioned

    C. fourier transform

    D. tools for integrating C/C++ and Fortran code

ANSWER: B

11. What is NumPy?

    A. Numeric Program

    B. Natural Python

    C. Numerical Python

    D. Nonlinear Python

ANSWER: C

12. Identify the default data type of NumPy arrays

    A. int32

    B. None of the above

    C. object

    D. float64

ANSWER: D

13. Identify which of the following is used to create an identity matrix in NumPy?

    A. eye()

    B. ones()

    C. arange()

    D. zeros()

ANSWER: A

14. Identify among the following which is used to find the maximum element in a NumPy array?

    A. max()

    B. amax()

    C. None

    D. Both A and B

ANSWER: D

15. Identify among the following, which is used to find the standard deviation of a NumPy array?

    A. None of the above

    B. var()

    C. Both A and B

    D. std()

ANSWER: D

16. Identify among the following which is used to find the sum of the elements in a NumPy array?

    A. All of the above

    B. mode()

    C. sum()

    D. None of the above

ANSWER: C

17. Identify among the following which is used to find the median of a NumPy array?

    A. mode()

    B. mean()

    C. median()

    D. None of the above

ANSWER: C

18. Identify among the following which is used to compute the dot product of two NumPy arrays?

    A. None of the above

    B. inner()

    C. All of the above

    D. dot()

ANSWER: D

19. Identify among the following which is used to find the inverse of a matrix in NumPy?

A. All of the above

B. reverse()

C. inv()

D. None of the above

ANSWER: C

20. Identify among the following libraries, which allows manipulating, transforming and visualizing data easily and efficiently.

A. All of the above

B. NumPy

C. Matplotlib

D. Pandas

ANSWER: A

21. What is PANDAS?

A. Panel Data

B. None

C. Panel Data Action

D. Panel Dashboard

ANSWER: A

22. Pandas Series can have what data types?

A. float

B. All of the above

C. String

D. integer

ANSWER: B

23. What is used when data is in Tabular Format?

A. NumPy

B. All of the above

C. Matplotlib

D. Pandas

ANSWER: D

24. What is a collection of data values and operations that can be applied to that data?

A. Table

B. Data Frame

C. Data Structure

D. None of the above

ANSWER: C

25. Identify among the following statement which is correct for importing pandas in python?

A. import pandas

B. All of the above

C. import pandas as pds

D. import pandas as pd

ANSWER: B

26. Identify among the following property/attribute which assign name to the Series?

A. size

B. index.name

C. name

D. Series.name

ANSWER: C

27. Identify among the following attributes that returns True if there is no value in Series?

A. empty

B. size

C. index

D. values

ANSWER: A

28. Identify among the following statement which shows first five values of Series 'S1'?

A. S1.head( )

B. S1.head(5)

C. None of the above

D. Both A and B

ANSWER: D

29. What we can perform on two series in Pandas?

A. Addition

B. All of the above

C. Multiplication

D. Subtraction

ANSWER: B

30. Identify the following method that is used to add two series?

A. add( )

B. addition( )

C. sum( )

D. None of the above

ANSWER: A

31. Identify among the following which fills the missing values in Series?

A. fill_value

B. fill-value

C. fill value

D. None of the above

ANSWER: A

32. Identify among the following function which is used for basic mathematical operations in Series?

A. add( )

B. All of the above

C. div( )

D. mul( )

ANSWER: B

33. Mathematical Operations on two Series object is done by matching what?

A. Both of the above

B. values

C. indexes

D. None of the above

ANSWER: C

34. Identify among the following statement which will return 10 values from the bottom/end of the Series 'S1'?

A. S1.tail( )

B. S1(10)

C. S1.head(10)

D. S1.tail(10)

ANSWER: D

35. Identify among the following which are valid operations on Series 'S1'?

A. >>> S1 + 2

B. All of the above

C. >>> S1 * 2

D. >>> S1 ** 2

ANSWER: B

36. Pandas is what kind of library?

A. Open source

B. Proprietary

C. Freeware

D. End source

ANSWER: A

37. We can generate what by using Matplotlib?

A. Histograms

B. All of the above

C. Scatterplots

D. Bar charts

ANSWER: B

38. Code a statement to import pandas with alias name 'pds'

A. import Pandas as pds

B. Import pandas as pds

C. none

D. import pandas as pds

ANSWER: D

39. By default Series have _____ data labels starting from ___.

A. numeric, zero

B. numeric, one

C. character, ‘a’

D. character, zero

ANSWER: A

40. Identify among the following which is parameter of Series( ) function.

A. data

B. All of the above

C. dtype

D. index

ANSWER: B

41. In Pandas what is used to store data in multiple columns?

A. Series

B. None of the above

C. Both of the above

D. DataFrame

ANSWER: D

42. What is a two-dimensional labelled data structure?

A. List

B. Series

C. DataFrame

D. None of the above

ANSWER: C

43. Identify the library that is to be imported for creating DataFrame

A. Pandas

B. DataFrame

C. Python

D. Random

ANSWER: A

44. Identify among the following function that is used to create DataFrame.

A. None of the Above

B. NewFrame( )

C. CreateDataFrame( )

D. DataFrame( )

ANSWER: D

45. The following code create a dataframe named ‘D1’ with how many columns?

```
import pandas as pd
D1 = pd.DataFrame([1,2,3] )
```

A. 3

B. 2

C. 1

D. 4

ANSWER: C

46. DataFrame created from single Series has how many column?

A. None of the above

B. 2

C. n (Where n is the number of elements in the Series)

D. 1

ANSWER: D

47. By default new column added as the _____ column in DataFrame.

A. Last (Right Side)

B. Second

C. First (Left Side)

D. Random

ANSWER: A

48. Identify the method which is used to delete row or column in DataFrame?

A. drop( )

B. del( )

C. delete( )

D. None of the above

ANSWER: A

49. The parameter axis of function drop( ) is assigned the value _____ to delete a row.

A. 2

B. 1

C. 0

D. 3

ANSWER: C

50. The parameter axis of function drop( ) is assigned the value _____ to delete a column.

A. 0

B. 3

C. 2

D. 1

ANSWER: D

51. Identify the method which is used to change the labels of rows and columns in DataFrame?

A. change( )

B. None of the above

C. replace( )

D. rename( )

ANSWER: D

52. Identify the value that should be given to axis parameter of rename function to alter column name?

A. rows

B. None of the above

C. index

D. columns

ANSWER: D

53. Identify the following parameter which is used to specify row or column in rename function of DataFrame?

A. rowindex

B. index

C. Both of the above

D. colindex

ANSWER: B

54. Following statement will display how many rows from DataFrame 'DF1'.

>>> DF1.head()

A. All

B. 5

C. 3

D. 2

ANSWER: B

55. Identify the following function which display the last 'n' rows from the DataFrame?

A. head( )

B. None of the above

C. end( )

D. tail( )

ANSWER: D

56. Identify the property of dataframe which is used to check that dataframe is empty or not?

A. empty

B. emp

C. isempty

D. None of the above

ANSWER: A

57. Identify the statement to display the row labels of dataframe ‘DF’.

A. DF.index

B. DF.indexrow( )

C. Index.DF

D. DF.row_index

ANSWER: A

58. How to display the first row of dataframe ‘DF’

A. print(DF.head(1))

B. All of the above

C. print(DF.iloc[0 : 1])

D. print(DF[0 : 1])

ANSWER: B

59. Identify the statement to display the data types of each column of dataframe ‘DF’.

A. DF.types( )

B. None of the above

C. dtypes.DF( )

D. DF.dtypes

ANSWER: D

60. Identify the statement to display the dimension of dataframe ‘DF’.

A. DF.dim

B. None of the above

C. DF.dim( )

D. DF.ndim

ANSWER: D

61. What means graphical or pictorial representation of the data using graph, chart, etc?

A. MatLab

B. Visual Data

C. Data visualization

D. None of the above

ANSWER: C

62. Identify among the following library which is to be imported for creating chart in python?

A. Math

B. Pandas

C. Matplotlib

D. Random

ANSWER: C

63. What function is used to display figure/chart?

A. showing( )

B. screen( )

C. display( )

D. show( )

ANSWER: D

64. Values which are displayed on x-axis is called _____

A. y ticks

B. None of the above

C. xy ticks

D. x ticks

ANSWER: D

65. plot(a, b) is provided with two parameters, which indicates values for what?

A. x-axis only

B. y-axis and x-axis, respectively

C. x-axis and y-axis, respectively

D. None of the above

ANSWER: C

66. Identify the attribute of plot( ) function which is used to set the edge color of bar in bar chart in matplotlib?

A. edgecolor

B. colorofedge

C. bordercolor

D. none of the above

ANSWER: A

67. Identify the attribute of plot( ) function which is used to set the different color of bars in bar chart in matplotlib?

A. none of the above

B. barcolor

C. colorbar

D. color

ANSWER: D

68. Identify how are the points in the domain set given as input to the algorithm?

A. Polynomials

B. Scalar points

C. Vector of features

D. Clusters

ANSWER: C

69. Identify what is not accessible to the learner?

A. Labeling Function

B. Label Set

C. Training Set

D. Domain Set

ANSWER: A

70. _____ is the error available to the learner.

A. training error

B. error of the classifier

C. true error

D. testing error

ANSWER: A

71. What is Machine learning?

A. The autonomous acquisition of knowledge through the use of computer programs

B. The selective acquisition of knowledge through the use of manual programs

C. The selective acquisition of knowledge through the use of computer programs

D. The autonomous acquisition of knowledge through the use of manual programs

ANSWER: A

72. Identify the key difference between supervised and unsupervised learning.

A. Supervised learning is always more accurate than unsupervised learning.

B. Supervised learning predicts labels, while unsupervised learning discovers patterns.

C. Supervised learning is used for classification, while unsupervised learning is used for regression.

D. Supervised learning requires labeled data, while unsupervised learning does not.

ANSWER: D

73. Identify the type of machine learning algorithm that falls under the category of “unsupervised learning”?

A. Linear Regression

B. Random Forest

C. Decision Trees

D. K-means Clustering

ANSWER: D

74. Identify the category where linear regression belong?

A. Neither supervised nor unsupervised learning

B. Supervised learning

C. Unsupervised learning

D. Both supervised and unsupervised learning

ANSWER: B

75. The learner is trying to predict housing prices based on the size of each house. Identify the type of regression here.

A. Linear Regression

B. Logistic Regression

C. Multivariate Logistic Regression

D. Multivariate Linear Regression

ANSWER: A

76. _____ minimizes the cost function.

A. Linear regression

B. Gradient descent

C. PAC learning

D. Polynomial regression

ANSWER: B

77. Identify the kind of algorithm about logistic regression?

A. Classification

B. Ranking

C. Regression

D. Cost function minimization

ANSWER: A

78. What is the output of training process in machine learning?

A. null

B. machine learning algorithm

C. machine learning model

D. accuracy

ANSWER: C

79. Identify among the following that is not Machine Learning?

A. artificial intelligence

B. none of the mentioned

C. both a and b

D. rule based inference

ANSWER: D

80. Identify among the following that is a characteristic of best machine learning method?

A. fast

B. all of the above

C. scalable

D. accuracy

ANSWER: B

81. In Machine Learning, what is Model Selection ?

A. Find interesting directions in data and find novel observations/ database cleaning

B. when a statistical model describes random error or noise instead of underlying relationship

C. The process of selecting models among different mathematical models, which are used to describe the same data set

D. All above

ANSWER: C

82. Between classifier predicted output and actual output, the average squared difference is called?

A. mean absolute error

B. root mean squared error

C. mean squared error

D. mean relative error

ANSWER: C

83. The _____ of the hyperplane depends upon the number of features.

A. reduction

B. classification

C. dimension

D. none of the above

ANSWER: C

84. Identify among the following, which is a categorical data?

A. weight of a person

B. expenditure in rupees

C. prize of house

D. branch of bank

ANSWER: D

85. Identify among the following which are supervised learning applications?

A. Bioinformatics, Speech recognition

B. Image classification, Real-time visual tracking

C. Autonomous car driving, Logistic optimization

D. Spam detection, Pattern detection, Natural Language Processing

ANSWER: D

86. Identify among the following which is a good test dataset characteristic?

A. large enough to yield meaningful results

B. is representative of the dataset as a whole

C. none of the above

D. both A and B

ANSWER: D

87. In linear model syntax (lm(formula,data,..)), data refers to _____

A. Matrix

B. List

C. Array

D. Vector

ANSWER: D

88. Many researchers in the last decade started training bigger and bigger models which are built with several different layers that's why this approach is called _____

A. Unsupervised learning

B. Machine learning

C. Reinforcement learning

D. Deep learning

ANSWER: D

89. What is the most general form of distance in ML?

A. manhattan

B. minkowski

C. mode

D. eucledian

ANSWER: D

90. Identify when it is necessary to allow the model to develop a generalization ability and avoid a common problem called _____

A. Classification

B. Overlearning

C. Overfitting

D. Regression

ANSWER: C

91. Identify what is true about Manhattan distance.

A. it can be used for categorical as well as continuous

B. it can be used for categorical variables

C. it can be used for continuous variables

D. it can be used for constants

ANSWER: C

92. What provides some built-in datasets that can be used for testing purposes in ML?

A. none of the above

B. classification

C. regression

D. scikit-learn

ANSWER: D

93. Which parameter allows specifying the percentage of elements to put into the test/training set?

A. All of the above

B. training_size

C. test_size

D. None of these

ANSWER: A

94. In a dataset how do you handle missing or corrupted data?

A. Drop missing rows or columns

B. All of the above

C. Assign a unique category to missing values

D. Replace missing values with mean/median/mode

ANSWER: B

95. What data is used to optimize the parameter settings of a supervised learner model?

A. training

B. validation

C. verification

D. test

ANSWER: B

96. Naive Bayes classifiers belong to which collection of algorithms?

A. Regression

B. Clustering

C. Classification

D. All of the above

ANSWER: C

97. Identify the techniques which involve the usage of both labeled and unlabeled data.

A. Supervised

B. None of the above

C. Unsupervised

D. Semi-supervised

ANSWER: D

98. Is it necessary that they have a linear relationship if two variables are correlated?

A. Yes

B. None of the above

C. Can't say

D. No

ANSWER: D

99. Identify among the following which is not supervised learning?

A. Naive Bayesian

B. Decision Tree

C. PCA

D. Linear regression

ANSWER: C

100. When some telecommunication company wants to categorize their customers into distinct groups ,this is an example of _____

A. unsupervised learning

B. reinforcement learning

C. supervised learning

D. data extraction

ANSWER: A

101. Identify what is 'Test set' in machine learning?

A. Training purpose

B. It is a set of data is used to discover the potentially predictive relationship.

C. Test set is used to test the accuracy of the hypotheses generated by the learner.

D. None of above

ANSWER: C

102. Identify among the following when scale data by removing elements that don't belong to a given range or by considering a maximum absolute value.

A. MinMaxScaler

B. MaxAbsScaler

C. None of the above

D. Both A and B

ANSWER: D

103. Identify the correct option - Attribute selection measures are also known as splitting rules.

A. FALSE

B. TRUE

C. Can’t say

D. None of the above

ANSWER: B

104. In _____ all labels are turned into sequential numbers.

A. DictVectorizer

B. LabelBinarizer class

C. LabelEncoder class

D. FeatureHasher

ANSWER: C

105. High entropy means that the partitions in classification are

A. pure

B. useless

C. useful

D. not pure

ANSWER: D

106. Number of classes must be _____ in multiclass classification

A. greater than two

B. equals to two

C. less than two

D. None of the above

ANSWER: A

107. Identify the purpose of performing cross-validation?

A. To assess the predictive performance of the models

B. To judge how the trained model performs outside the sample on test data

C. None of these

D. Both A and B

ANSWER: D

108. In a data-set, a measurable property or parameter is?

A. training data

B. validation data

C. test data

D. feature

ANSWER: D

109. Bernoulli Nave Bayes Classifier is _____ distribution

A. Binary

B. Discrete

C. Continuous

D. none of these

ANSWER: A

110. Identify the correct option - Conditional probability is a measure of the probability of an event given that another event has already occurred.

A. None of the above

B. FALSE

C. Can’t say

D. TRUE

ANSWER: D

111. Given a seismic data and you want to predict next earthquake , this is an example of

A. dimensionality reduction

B. reinforcement learning

C. unsupervised learning

D. supervised learning

ANSWER: D

112. Identify the correct statement about 'Training set'?

A. Training set is used to test the accuracy of the hypotheses generated by the learner.

B. None of above

C. Both A & B

D. A set of data is used to discover the potentially predictive relationship.

ANSWER: D

113. Identify the correct option - Overfitting is more likely when you have huge amount of data to train?

A. TRUE

B. None of the above

C. Can't say

D. FALSE

ANSWER: D

114. Gaussian Naive Bayes Classifier is _____ distribution

A. Binary

B. Discrete

C. Continuous

D. None of these

ANSWER: C

115. Identify among these which is a tree based learner?

A. random forest

B. bayesian belief network

C. bayesian classifier

D. rule based

ANSWER: A

116. Identify among the following which is true about Residuals?

A. Depend on the situation

B. Higher is better

C. Lower is better

D. None of these

ANSWER: C

117. What is the type of dataset available in Supervised Learning?

A. unlabeled dataset

B. excel file

C. csv file

D. labeled dataset

ANSWER: D

118. What is another name for an output attribute.

A. predictive variable

B. dependent variable

C. estimated variable

D. independent variable

ANSWER: D

119. What do you mean by learning?

A. learning is the ability to change according to external stimuli and remembering most of all previous experiences.

B. a set of data is used to discover the potentially predictive relationship.

C. robots are programed so that they can perform the task based on data they gather from sensors.

D. it is a set of data is used to discover the potentially predictive relationship.

ANSWER: A

120. The problem of finding hidden structure in unlabeled data is called _____

A. supervised learning

B. none of the above

C. reinforcement learning

D. unsupervised learning

ANSWER: D

121. Identify which library is used to create statistical graphics in Python?

A. Seaborn

B. Numpy

C. Pandas

D. scikit-learn

ANSWER: A

122. Identify which of the following is NOT a type of plot available in Seaborn?

A. Line plot

B. Tree plot

C. Bar plot

D. Scatter plot

ANSWER: B

123. Identify the function which is used to create a histogram in Seaborn? # import seaborn as sns

A. sns.histplot()

B. sns.scatterplot()

C. sns.lineplot()

D. sns.barplot()

ANSWER: A

124. Identify the function which is used to create a heatmap in Seaborn? # import seaborn as sns

A. sns.barplot()

B. sns.scatterplot()

C. sns.lineplot()

D. sns.heatmap()

ANSWER: D

125. Identify the function which is used to create a box plot in Seaborn? # import seaborn as sns

A. sns.lineplot()

B. sns.scatterplot()

C. sns.boxplot()

D. sns.barplot()

ANSWER: C

126. Identify the function, which is used to create a violin plot in Seaborn? # import seaborn as sns

A. sns.scatterplot()

B. sns.violinplot()

C. sns.lineplot()

D. sns.barplot()

ANSWER: B

127. Identify the function which is used to create a scatter plot in Seaborn? # import seaborn as sns

A. sns.violinplot()

B. sns.boxplot()

C. sns.scatterplot()

D. sns.lineplot()

ANSWER: C

128. Identify the function which is used to create a line plot in Seaborn? # import seaborn as sns

A. sns.scatterplot()

B. sns.boxplot()

C. sns.violinplot()

D. sns.lineplot()

ANSWER: D

129. What is Seaborn?

A. All of the above

B. provides a high-level interface for creating visually appealing and informative statistical graphics.

C. Built on top of the popular Matplotlib library.

D. open-source Python library designed for data visualization

ANSWER: A

130. Choose the right way to install Seaborn?

A. seaborn install pip

B. pip install seaborn

C. Both A and B

D. None of the above

ANSWER: B

131. Identify some industry applications of Seaborn?

A. Finance

B. All of the above

C. Marketing

D. Healthcare

ANSWER: B

132. Identify among the following which is not a function in excel?

A. SUBTRACT

B. MIN

C. SUM

D. MAX

ANSWER: A

133. Identify the main components of big data?

A. HDFS

B. All of the above

C. YARN

D. MapReduce

ANSWER: B

134. Identify the use of data cleaning?

A. To remove the noisy data

B. All of the above

C. Correct the inconsistencies in data

D. Transformations to correct the wrong data

ANSWER: B

135. Identify among the following which are the Benefits of Big Data Processing?

A. Businesses can utilize outside intelligence while taking decisions.

B. All of the above

C. Improve customer service

D. Better operational efficiency

ANSWER: B

136. Identify among the following which is not a part of the data science process?

A. Model planning

B. Discovery

C. Operationalize

D. Communication building

ANSWER: D

137. Data in ____ bytes size is called big data

A. Meta

B. Peta

C. Tera

D. Giga

ANSWER: B

138. When users submit incorrect data in the compulsory fields to avoid divulging personal information. Then it is called _____

A. Characteristic missing data

B. Organized missing data

C. Disguised missing data

D. Coordinated missing data

ANSWER: C

139. Identify among the following which is not true about data reduction?

A. Reduced data strives to gives same analytical results as the original data

B. It involves numerosity reduction

C. It involves dimensionality reduction

D. Reduced data gives strives to give less accurate analytical results the original data

ANSWER: D

140. Identify among the following which is an appropriate measure of central tendency to use fill in missing values of an attribute of a skewed distribution?

A. Median

B. Weighted mean

C. Mean

D. Geometric mean

ANSWER: A

141. Identify among the following which is the most effective measure of the center of symmetric data set?

A. Mean

B. Midrange

C. Mode

D. Median

ANSWER: A

142. Identify among the following which is not true about scatter plots?

A. It is used to identify outliers

B. It is used to identify relationship between attributes

C. It is used to identify clusters

D. It is used in the case of univariate distribution

ANSWER: D

143. Identify among the following which is not a proximity measure?

A. Probability measures

B. Similarity measures

C. Dissimilarity measures

D. Distance measures

ANSWER: A

144. Identify among the following technique which predict data values using results derived from different data?

A. Association rules

B. Summarization

C. Sequence discovery

D. Regression

ANSWER: D

145. Identify among the following which is based on predictive modeling?

A. Sequence discovery

B. Summarization

C. Time series analysis

D. Clustering

ANSWER: C

146. The analysis of the data points that deviate from the general expected behavior of data in the data set is called _____

A. Anomaly analysis

B. Relevance analysis

C. Cluster analysis

D. Regression analysis

ANSWER: A

147. The data is stored in _____ after cleaning and integrating data from heterogeneous sources.

A. Data Warehouse

B. Database

C. Flat files

D. Directories

ANSWER: A

148. There are many ways in which various types of data can be stored. Choose the attribute which is not relevant to store data.

A. Executable Attributes

B. Nominal Attributes

C. Binary Attributes

D. Ordinal Attributes

ANSWER: A

149. Identify among the following which is not the name for the data whose values are simply names?

A. Symmetrical data

B. Categorical Attributes

C. Nominal Attributes

D. Enumerations

ANSWER: A

150. Identify among the following which can be appropriately represented as an ordinal attribute?

A. Gender of a student

B. Probability of rain

C. Hair color

D. Ratings in a survey

ANSWER: D

151. The attribute type with values that are measurable quantities represented by integers or real values is called as _____

A. Numeric

B. Binary

C. Nominal

D. Ordinal

ANSWER: A

152. Identify among the following function which stacks 1D arrays as columns into a 2D array?

A. row_stack

B. all of the mentioned

C. com_stack

D. column_stack

ANSWER: D

153. Identify among the following method which creates a new array object that looks at the same data?

A. paste

B. copy

C. view

D. all of the mentioned

ANSWER: C

154. Evaluate the result of the following code?

```
import numpy as np
a = np.array([10, 20, 30])
b = np.array([40, 50, 60])
c = np.outer(a, b)
print(c)
```

A. [[1, 2, 3], [4, 5, 6]]

B. [[400 500 600],[800 1000 1200],[1200 1500 1800]]

C. [1, 2, 3, 4, 5, 6]

D. Error

ANSWER: B

155. Evaluate the result of the following code?

```
import numpy as np
a = np.array([11, 21, 31])
b = np.array([41, 51, 61])
c = np.dot(a, b)
print(c)
```

A. [[1, 2, 3], [4, 5, 6]]

B. [[1, 4], [2, 5], [3, 6]]

C. [1, 2, 3, 4, 5, 6]

D. 3413

ANSWER: D

156. Identify among the following statement which will create an empty series named “S1”?

```
# import pandas as pd
```

A. S1 = pd.Series(None)

B. S1 = pd.Series( )

C. None of the above

D. Both of the above

ANSWER: D

157. Identify the number of elements that will be there in the series named “S1”? # import pandas as pd

```
>>> S1 = pd.Series(range(7))
>>> print(S1)
```

A. 7

B. 4

C. 6

D. None of the above

ANSWER: A

158. The keys of dictionary become _____ when we create a series from dictionary

A. Caption of the series

B. Value of the series

C. Index of the series

D. None of the series

ANSWER: C

159. Choose the value to get the ouput as 3

```
import pandas as pnd
S1=pnd.Series([1,2,3,4], index = ['a','b','c','d'])
print(S1[_____])
```

A. ‘c’

B. 2

C. None of the above

D. Both A and B

ANSWER: D

160. How many values will be modified by last statement of given code ?

```
import pandas as pd
S1 = pd.Series(['NewDelhi', 'WashingtonDC', 'London', 'Paris'],
index=['A', 'B', 'C', 'D'])
S1['A' : 'C'] = 'ND'
```

A. 3

B. 2

C. 1

D. 4

ANSWER: A

161. S1.values will return all the values of Series ‘S1’ in what?

A. List

B. Dictionary

C. Tuple

D. String

ANSWER: A

162. Identify among the following property/attribute that return total number of values in Series ‘S1’?

A. index

B. values

C. size

D. None of the above

ANSWER: C

163. Identify among the following attributes that returns all the values of Series?

A. size

B. values

C. name

D. index

ANSWER: B

164. Evaluate the result of the following code:

```
import pandas as pd
S1=pd.Series()
print(pd.Series().empty)
```

A. False

B. True

C. Error

D. None of the above

ANSWER: B

165. Evaluate the result of the following code :

```
import pandas as pd
S1=pd.Series([11,21,31,41])
S2=pd.Series([71,81])
S3=S1+S2
print(S3.size)
```

A. 2

B. 4

C. 6

D. Error

ANSWER: B

166. Evaluate the result of the following :

```
import pandas as pd
S1=pd.Series([11,21,31,41])
S2=pd.Series([71,81])
print((S1+S2).count())
```

A. 6

B. 4

C. 2

D. 0

ANSWER: C

167. Identify among the following that returns number of non-NaN values of Series?

A. index

B. size

C. count

D. values

ANSWER: C

168. Evaluate the result of the following :

```
import pandas as pd
S1=pd.Series([11,21,31,41])
S2=pd.Series([71,81,91,101])
S2.index=['a','b','c','d']
print((S1+S2).count())
```

A. 8

B. 4

C. 0

D. 6

ANSWER: C

169. When an operation is carried out on every value of Series object is called _____

A. Scalar Operation

B. None of the above

C. Both of the above

D. Vector Operation

ANSWER: D

170. In given code S2 is _____ ?

S2 = S1[2 : 5] #S1 is a Series object

A. Series

B. Tuple

C. List

D. None of the above

ANSWER: A

171. Pandas Series is size _____ and value _____

A. Immutable, Mutable

B. Immutable, Immutable

C. Mutable, Mutable

D. Mutable, Immutable

ANSWER: A

172. Pandas DataFrame is size _____ and value _____

A. Mutable, Immutable

B. Immutable, Immutable

C. Immutable, Mutable

D. Mutable, Mutable

ANSWER: D

173. Identify the data type of given series 'S1'?

import pandas as pd

S1=pd.Series('a', index=(12.0, 13.0, 14.0, 15.0))

print(S1)

A. float64

B. int64

C. object

D. string

ANSWER: C

174. In given series 'S1', how many elements will be there?

import pandas as pd

S1=pd.Series('python practice')

A. 0

B. 15

C. 2

D. 1

ANSWER: D

175. Two common ways for accessing the elements of a series are _____ and _____

A. Indexing, Slicing

B. Labelled Indexing, Positional Indexing

C. Indexing, Concatenation

D. Slicing, Cutting

ANSWER: A

176. Identify the following statement which is correct to add NaN value in series?

A. S1=pd.Series([10, np.NaN,11])

B. S1=pd.Series([10, None, 11])

C. None of the above

D. Both of the above

ANSWER: D

177. Identify the following attribute of Series that returns the tuple?

A. size

B. index

C. values

D. shape

ANSWER: D

178. Identify the type of error that is returned, when the length of index and the length of data in Series() function is not same?

A. Key Error

B. Name Error

C. Syntax Error

D. Value Error

ANSWER: D

179. Choose the statement to display 12.5 as output using positional indexing.

```
import pandas as pd
S1=pd.Series([11, 12.5, None, 6],
index=["J","F","M","A"])
```

A. print(S1[1])

B. print(S1[“F”])

C. print(S1[0])

D. print(S1[“J”])

ANSWER: B

180. Can a Series have duplicate index value?

A. No

B. Yes

C. Yes, Only series with integer values

D. Yes, Only series with character values

ANSWER: B

181. David wants to display the series ‘S1’ in reverse order. How to do that?

A. >>> S1[: : 1]

B. >>> S1[: 1 1]

C. >>> S1[-1 : :]

D. >>> S1[: : -1]

ANSWER: D

182. Number of students in each section of class 9th is stored in series ‘P1’. Write a statement to change the value of section ‘A’ and ‘B’ to 50.

A 41

B 44

C 40

D 42

E 47

A. P1[: 2] = 50

B. P1[0 : 2] = 50

C. None of the above

D. Both of the above

ANSWER: D

183. Series ‘S1’ has five values with index value (0, 1, 2, 3, 4) and series ‘S2’ has five values with index (2, 3, 4, 5, 6). What will be the total number of values in ‘S3’ if S3 = S1 + S2

A. 5

B. 6

C. 8

D. 7

ANSWER: D

184. print(P1[-1]) will return _____ #’P1′ is a series

A. last element of series ‘P1’

B. first element of series ‘P1’

C. all elements of series ‘P1’

D. Key Error

ANSWER: D

185. >>> P1[1:3] = 50 will update the value of _____ elements

A. 1

B. 4

C. 3

D. 2

ANSWER: D

186. Evaluate the result of the following:

```
import pandas as pd
S1=pd.Series(data=[11, 12, None, 6,9,7],index=[1,12,3,4,2,4])
print(S1.count())
```

A. 5

B. 6

C. 4

D. Error

ANSWER: A

187. Index matching is implemented and all missing values are filled in with _____ by default while performing mathematical operations on series.

A. 1

B. None

C. 0

D. NaN

ANSWER: D

188. Identify the following which is used to give user defined column index in DataFrame?

A. columns

B. column

C. index

D. colindex

ANSWER: A

189. The following code create a dataframe named 'P1' with _____ rows.

```
import pandas as pd
y = [{'a':10, 'b':20}, {'a':5, 'b':10, 'c':20}]
P1 = pd.DataFrame(y)
```

A. 2

B. 1

C. 0

D. 3

ANSWER: A

190. Dictionary keys will become _____ when we create DataFrame from List of Dictionaries.

A. None of the above

B. Row labels

C. Both of the above

D. Column labels

ANSWER: D

191. Number of columns in DataFrame is equal to the _____ when we create DataFrame from List of Dictionaries.

A. maximum number of keys in first dictionary of the list

B. None of the above

C. maximum number of dictionaries in the list

D. maximum number of different keys in all dictionaries of the list

ANSWER: D

192. Number of rows in DataFrame is equal to the _____ when we create DataFrame from List of Dictionaries.

A. number of dictionaries in the list

B. maximum number of keys in any dictionary of the list

C. maximum number of keys in first dictionary of the list

D. None of the above

ANSWER: A

193. In the following code, if column 'Rollno' already exists in the DataFrame 'P1' then the assignment statement will _____

```
P1['Rollno'] = [1,2,3] #There are only three rows in DataFrame P1'
```

A. Return error

B. None of the above

C. Add new column

D. Replace the already existing values.

ANSWER: D

194. PF1.loc[ ] method is used to _____ # PF1 is a DataFrame

A. Add new row in a DataFrame 'PF1'

B. To change the data values of a row to a particular value

C. None of the above

D. Both of the above

ANSWER: D

195. If the DataFrame has more than one row with the same label, then DataFrame.drop( ) method will delete _____

A. first matching row from it.

B. Return Error.

C. last matching row from it.

D. all the matching rows from it.

ANSWER: D

196. In a dataframe, the parameter axis='index' of rename( ) function is used to specify that the _____

A. row label is to be changed

B. column label is to be changed

C. row and column label is to be changed

D. None of the above.

ANSWER: A

197. What will happen if in the rename( ) function we pass only a value for a row label that does not exist?

A. the existing row label will be left as it is.

B. matching row label will not change.

C. it returns an error.

D. None of the above

ANSWER: A

198. The following statement will return the column as a _____

>>> DF.loc[: , 'Name'] #DF is a DataFrame object

A. DataFrame

B. Tuple

C. List

D. Series

ANSWER: D

199. Choose the correct statement?

dF1=dF1.append(dF2) #dF1 and dF2 are DataFrame object

A. We are appending dF1 in dF2

B. None of the above

C. We are creating Series from DataFrame

D. We are appending dF2 in dF1

ANSWER: D

200. What parameter is used in append( ) function of DataFrame to get the column labels in sorted order?

A. sort

B. sorter

C. sorted

D. None of the above

ANSWER: A

201. The append() method of DataFrame can also be used to append _____to a DataFrame

A. Series

B. Dictionary

C. None of the above

D. Both of the above

ANSWER: D

202. Identify among the following attribute of DataFrame which is used to display row labels?

A. columns

B. values

C. dtypes

D. index

ANSWER: D

203. Identify among the following attribute of DataFrame which is used to display data type of each column in DataFrame?

A. dtypes

B. Typesdata

C. types

D. datatypes

ANSWER: A

204. It means _____ if the following statement return (7, 4)

>>> DF.shape #DF is a DataFrame object

A. DataFrame DF has 4 rows 5 columns

B. DataFrame DF has 7 rows 4 columns

C. DataFrame DF has 4 rows 5 rowlabels

D. None of the above

ANSWER: B

205. We can write _____ to plot x versus y #plt is an alias for matplotlib.pyplot

A. plt.plot(y, x)

B. plt.plot(x,y)

C. plt.plot(x)

D. None of the above

ANSWER: B

206. Write a code to display “population” as x-axis label. (consider plt as an alias name of matplotlib.pyplot)

A. plt.label(“population”)

B. plt.xlabel(“population”)

C. plt.xlabel(population)

D. None of the above

ANSWER: B

207. Attribute/parameter to set marker size is _____

A. markersize

B. sizeofmarker

C. sizemarker

D. None of the above

ANSWER: A

208. _____ are column-charts, where each column represents a range of values, and the height of a column corresponds to how many values are in that range.

A. Bar graph

B. pie chart

C. Line chart

D. Histograms

ANSWER: D

209. Identify the parameter of plot( ) function which help to set the values of bins in Histogram?

A. bincount

B. none of the above

C. binvalue

D. bins

ANSWER: D

210. If we want to plot a line chart for values of list ‘a’ vs values of list ‘b’.Fill in the blank in the given code,

a = [11, 21, 31, 41, 51]

b = [10, 20, 30, 40, 50]

import matplotlib.pyplot as plt

plt.plot _____

A. [a, b]

B. (b, a)

C. (a, b)

D. None of the above

ANSWER: C

211. What the following code will create?

import matplotlib.pyplot as pl

a = [10,20,30,40,50]

b = [100, 200, 300, 400, 500]

c = [50, 100, 150, 200, 250]

pl.plot(a,b)

pl.plot(a,c)

pl.show()

A. line chart

B. bargraph

C. histogram

D. None of the above

ANSWER: A

212. The following code will show how many figure/chart?

import matplotlib.pyplot as pl

a = [10,20,30,40,50]

b = [100, 200, 300, 400, 500]

c = [50, 100, 150, 200, 250]

pl.plot(a,b)

pl.plot(a,c)

pl.show()

A. 1

B. 2

C. 3

D. 4

ANSWER: B

213. Choose among the options about how do we learn concepts from training examples?

A. Incrementally

B. Decremental

C. Arbitrarily

D. Non-incremental

ANSWER: A

214. The learner is trying to predict housing prices based on the size of each house. What kind of variable “size” is?

A. dependent variable

B. independent variable

C. label set variable

D. target variable

ANSWER: B

215. The target variable is represented along which axis?

A. Depends on the dataset

B. X axis

C. Either Y-axis or X-axis, it doesn’t matter

D. Y axis

ANSWER: D

216. The independent variable is represented along which axis?

A. Either X-axis or Y-axis, it doesn't matter

B. X axis

C. Y axis

D. Depends on the dataset

ANSWER: B

217. Can a cancer detection problem be solved by logistic regression?

A. Sometimes

B. Yes

C. No

D. Depends on the dataset

ANSWER: B

218. In machine learning what characterize unlabeled examples?

A. there is prior knowledge

B. there is no confusing knowledge

C. there is no prior knowledge

D. there is plenty of confusing knowledge

ANSWER: C

219. Identify from the following which is true about Naive Bayes?

A. Assumes that all the features in a dataset are equally important

B. Assumes that all the features in a dataset are independent

C. None of the above option

D. Both A and B

ANSWER: D

220. Identify among the following statements which best describes our approach to learning decision trees

A. identify the best approximation of the above by the greedy approach (to identifying the partitions)

B. identify the best partition of the input space and response per partition to minimise sum of squares error

C. identify the model which gives the best performance using the greedy approximation (option (b)) with the smallest partition scheme

D. identify the model which gives performance close to the best greedy approximation performance (option (b)) with the smallest partition scheme

ANSWER: A

221. Which technique associates a conditional probability value with each data instance.

A. linear regression

B. simple regression

C. logistic regression

D. multiple linear regression

ANSWER: C

222. Logistic regression is a ____ regression technique that is used to model data having a _____ outcome.

A. linear, numeric

B. nonlinear, binary

C. nonlinear, numeric

D. linear, binary

ANSWER: B

223. You are given reviews of few netflix series marked as positive, negative and neutral. Classifying reviews of a new netflix series is an example of

A. reinforcement learning

B. unsupervised learning

C. semisupervised learning

D. supervised learning

ANSWER: D

224. In Navie Bayes Classifier features being classified is _____ of each other

A. partial dependent

B. dependent

C. independent

D. none

ANSWER: C

225. Identify among the following option which is true about k-NN algorithm?

A. it can be used in both classification and regression

B. it can be used for regression

C. it can be used for classification

D. not useful in ml algorithm

ANSWER: A

226. How feature can be used?

A. binary split

B. predictor

C. none of the above

D. both a and b

ANSWER: D

227. Regression trees are often used to model _____ data.

A. linear

B. symmetrical

C. categorical

D. nonlinear

ANSWER: D

228. Choose the right option about the leaf nodes of a model tree.

A. linear regression equations.

B. nonlinear regression equations.

C. averages of numeric output attribute values.

D. sums of numeric output attribute values.

ANSWER: A

229. Choose how to select best hyper parameters in tree based models?

A. measure performance over training data

B. random selection of hyper parameters

C. both of these

D. measure performance over validation data

ANSWER: D

230. Which would you address using a supervised learning Algorithm of the Following Examples?

A. find the patterns in market basket analysis

B. given a set of news articles found on the web, group them into set of articles about the same story.

C. given a database of customer data, automatically discover market segments and group customers into different market segments.

D. given email labeled as spam or not spam, learn a spam filter

ANSWER: D

231. _____ distribution is a Multinomial Nave Bayes Classifier

A. continuous

B. None of these

C. binary

D. discrete

ANSWER: D

232. Identify among the following statements about Naive Bayes which is incorrect?

A. Attributes are equally important.

B. Attributes can be nominal or numeric

C. Attributes are statistically independent of one another given the class value.

D. Attributes are statistically dependent of one another given the class value.

ANSWER: D

233. Identify which one of these is not a tree based learner?

A. bayesian classifier

B. id3

C. cart

D. random forest

ANSWER: A

234. What is gini index?

A. it is a type of index structure

B. None of the options

C. both a and b

D. it is a measure of purity

ANSWER: D

235. Multivariate split is where the partitioning of tuples is based on a combination of attributes rather than on a single attribute.

A. None of the above

B. FALSE

C. Can't say

D. TRUE

ANSWER: D

236. Identify the standard approach to supervised learning?

A. a set of observed instances tries to induce a general rule

B. group the set of example into the training set and the test

C. split the set of example into the training set and the test

D. learns programs from data

ANSWER: C

237. Identify the option which can accept a NumPy RandomState generator or an integer seed.

A. make_blobs

B. training_size

C. test_size

D. random_state

ANSWER: D

238. Identify among the following step / assumption in regression modeling that impacts the trade-off between under-fitting and over-fitting the most.

A. The use of a constant-term

B. Whether we learn the weights by matrix inversion or gradient descent

C. The polynomial degree

D. None of these

ANSWER: C

239. Choose on how can you avoid overfitting?

A. by using validation only

B. by using inductive machine learning

C. by using a lot of data

D. None of above

ANSWER: C

240. Tree/Rule based classification algorithms generate _____ rule to perform the classification.

A. do while

B. while.

C. if-then.

D. switch.

ANSWER: C

241. In the regression equation Y = 45.23 + 0.50X, the intercept is

A. 0.5

B. 45.23

C. 1

D. indeterminable

ANSWER: B

242. Selecting data so as to assure that each class is properly represented in both the training and test set.

A. cross validation

B. bootstrapping

C. verification

D. stratification

ANSWER: D

243. What produce sparse matrices of real numbers that can be fed into any machine learning model?

A. DictVectorizer

B. FeatureHasher

C. None of the above

D. Both A and B

ANSWER: D

244. Identify the measure of goodness of fit for the estimated regression equation.

A. multiple coefficient of determination

B. mean square due to regression

C. mean square due to error

D. None of the above

ANSWER: B

245. Identify among the following that is true about averaging ensemble?

A. it can be used in both classification as well as regression

B. it can only be used in regression problem

C. it can only be used in classification problem

D. none of these

ANSWER: A

246. Dimensionality Reduction Algorithms are one of the possible ways to reduce the computation time required to build a model

A. None of the above

B. FALSE

C. Can't say

D. TRUE

ANSWER: D

247. Choose from the options on how SVM can be classified?

A. it is a model trained using supervised learning. it can be used for classification and regression.

B. it is a model trained using unsupervised learning. it can be used for classification but not for regression.

C. it is a model trained using unsupervised learning. it can be used for classification and regression.

D. t is a model trained using unsupervised learning. it can be used for classification but not for regression.

ANSWER: A

248. Linear SVMs have no hyper parameters that need to be set by cross-validation

A. FALSE

B. TRUE

C. Can't say

D. None of the above

ANSWER: A

249. Identify among the following which are components of generalization Error?

A. bias

B. variance

C. None of them

D. Both of them

ANSWER: D

250. Identify among the following sentence which is correct?

A. machine learning relates with the study, design and development of the algorithms that give computers the capability to learn without being explicitly programmed.

B. data mining can be defined as the process in which the unstructured data tries to extract knowledge or unknown interesting patterns.

C. None of the above

D. Both A and B

ANSWER: D

251. Choose the right option on the effectiveness of an SVM that depends upon:

A. Selection of Kernel

B. All of the above

C. Soft Margin Parameter C

D. Kernel Parameters

ANSWER: B

252. Which of the following plot best suited to test linear relationship of y(dependent) and x(independent) continuous variables

A. Histograms

B. Bar chart

C. Scatter plot

D. None of these

ANSWER: C

253. SVM algorithms use a set of mathematical functions that are defined as the kernel.

A. FALSE

B. TRUE

C. Can’t say

D. None of the above

ANSWER: B

254. Choose the right option on tree based classifiers?

A. classifiers which form a tree with each attribute at one level

B. classifiers which perform series of condition checking with one attribute at a time

C. None of the above

D. Both A and B

ANSWER: D

255. Choose the right option on the "curse of dimensionality" which refers _____

A. all the problems that arise when working with data in the higher dimensions, that did not exist in the higher dimensions.

B. all the problems that arise when working with data in the lower dimensions, that did not exist in the higher dimensions.

C. all the problems that arise when working with data in the lower dimensions, that did not exist in the lower dimensions.

D. all the problems that arise when working with data in the higher dimensions, that did not exist in the lower dimensions.

ANSWER: A

256. scikit-learn also provides functions for creating dummy datasets from scratch:

A. make_classification()

B. all of the above

C. make_blobs()

D. make_regression()

ANSWER: B

257. Support vectors are the data points that lie closest to the decision surface.

A. FALSE

B. TRUE

C. Can’t say

D. None of the above

ANSWER: B

258. What is the term used to describe the case when the independent variables in a multiple regression model are correlated?

A. multicollinearity

B. correlation

C. regression

D. None of the above

ANSWER: A

259. Identify the two steps of tree pruning work?

A. pessimistic pruning and optimistic pruning

B. None of the above

C. cost complexity pruning and time complexity pruning

D. postpruning and prepruning

ANSWER: D

260. The difference between the actual Y value and the predicted Y value in a regression equation is called the _____

A. slope

B. scatter plot

C. outlier

D. residual

ANSWER: D

261. The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space(N the number of features) that distinctly classifies the data points.

A. FALSE

B. TRUE

C. Can’t say

D. None of the above

ANSWER: B

262. In many classification problems, the target _____ is made up of categorical labels which cannot immediately be processed by any algorithm.

A. random_state

B. All of the above

C. test_size

D. dataset

ANSWER: D

263. Identify among the following which is the main reason for pruning a Decision Tree?

A. to avoid overfitting the training set

B. to save space for storing the decision tree

C. to make the training set error smaller

D. to save computing time during testing

ANSWER: A

264. Identify the function of 'Supervised Learning'?

A. classifications, predict time series, annotate strings

B. speech recognition, regression

C. None of above

D. Both A and B

ANSWER: D

265. Identify among the following algorithm which comes under the classification?

A. k-nearest neighbor

B. brute force

C. dbscan

D. apriori

ANSWER: A

266. Point out the differences between Seaborn and Matplotlib

A. Seaborn provides a higher-level interface for creating visualizations.

B. All of the above

C. Seaborn has better integration with the Pandas library

D. Seaborn offers several specialized plots designed for statistical analysis.

ANSWER: B

267. Point out the advantages of using Seaborn

A. Seaborn simplifies the process of creating complex visualizations

B. All of the above

C. Seaborn offers a variety of specialized statistical plots.

D. It helps create attractive and professional-looking plots with minimal effort.

ANSWER: B

268. Identify the limitations of Seaborn

A. It may not provide as much control over customization as Matplotlib.

B. Seaborn can be slower than Matplotlib for certain types of plots.

C. All of the above

D. Both A and B

ANSWER: D

269. How to set Seaborn color palettes? # import seaborn as sns

A. None of the above

B. color_palette("palette_name")

C. Both A and B

D. sns.set_palette("palette_name")

ANSWER: D

270. Point out the basic Seaborn functions

A. load_dataset()

B. All of the above

C. jointplot()

D. distplot()

ANSWER: B

271. How to add labels, annotations to Seaborn plots?

A. using text() function

B. using annotate() function

C. None of the above

D. Both A and B

ANSWER: D

272. Choose the correct option on how categorical data are used in Seaborn visualizations?

A. boxplot()

B. All of the above

C. swarmplot()

D. violinplot()

ANSWER: B

273. Choose the correct option on how are numerical data used in Seaborn visualizations?

A. distplot()

B. kdeplot()

C. All of the above

D. jointplot()

ANSWER: C

274. Seaborn be integrated with what data visualization libraries?

A. scikit-learn

B. Pandas

C. Plotly

D. None of the above

ANSWER: C

275. Identify examples of Seaborn usage in data analysis and interpretation

A. Exploratory data analysis

B. Feature selection

C. None of the above

D. Both A and B

ANSWER: D

276. Identify the best practices for using Seaborn in data visualization projects?

A. Choose appropriate plot types

B. All of the above

C. Maintain readability

D. Use color effectively

ANSWER: B

277. Identify an MS-Excel feature that displays only the data in column (s) according to specified criteria?

A. Formula

B. Sorting

C. Filtering

D. Pivot

ANSWER: C

278. Identify the programming language used to write a Macro in MS Excel?

A. Java

B. C+

C. Visual basic

D. C++

ANSWER: C

279. Identify among the following, which is true about big data?

A. Big data can be processed using traditional techniques

B. Big data has low velocity meaning that it is generated slowly

C. Big data analysis does not involve reporting and data mining techniques

D. Big data refers to data sets that are at least a petabyte in size

ANSWER: D

280. Identify main components of Big Data.

A. HDFS

B. All of the above

C. YARN

D. MapReduce

ANSWER: B

281. What is a collection of data that is used in volume, yet growing exponentially with time.

A. Big Database

B. Big Data

C. Big Datafile

D. Big DBMS

ANSWER: B

282. Identify the primary characteristics of big data among the following

A. Value

B. Variety

C. All of the above

D. Volume

ANSWER: C

283. DSS in data warehouse stands for _____

A. Decision single system

B. Data storable system

C. Data support system

D. Decision support system

ANSWER: D

284. Identify the different features of Big Data Analytics.

A. Open-source

B. All of the above

C. Scalability

D. Data recovery

ANSWER: B

285. Identify the incorrect big data Technologies.

A. Apache Hadoop

B. Apache Kafka

C. Apache Pytorch

D. Apache Spark

ANSWER: C

286. During which phase of the data life cycle, a business decides what kind of data it needs, how it will be managed, who will be responsible for it, and the optimal outcomes.

A. archive

B. capture

C. manage

D. planning

ANSWER: D

287. In the data life cycle, which phase involves gathering data from various sources and bringing it into the organization?

A. Capture

B. Analyze

C. Archive

D. Manage

ANSWER: A

288. A data analyst finishes using a dataset, so they erase or shred the files in order to protect private information. This is called archiving.

A. True

B. None of the above

C. Can't say

D. False

ANSWER: D

289. A dairy farmer decides to open an ice cream shop on her farm. After surveying the local community about people's favorite flavors, she takes the data they provided and stores it in a secure hard drive so it can be maintained safely on her computer. This is part of which phase of the data life cycle?

A. Analyze

B. Plan

C. Manage

D. Archive

ANSWER: C

290. After opening the ice cream shop on her farm, the dairy farmer surveys the local community about people's favorite flavors. She uses the data she collected to determine that the top five flavors are strawberry, vanilla, chocolate, mint chip, and peanut butter. She feels confident in her decision to sell these flavors. This is part of which phase of the data life cycle?

A. Analyze

B. Plan

C. Capture

D. Archive

ANSWER: A

291. The data analysis process phases are ask, prepare, process, analyze, share, and act. What do data analysts do during the ask phase?

A. Clean the data

B. Create data visualizations

C. Collect and store data

D. Define the problem to be solved

ANSWER: D

292. The data analysis process phases are ask, prepare, process, analyze, share, and act. A data analyst cleans data to ensure it's complete and correct during the process phase.

A. None of the above

B. False

C. Can't say

D. True

ANSWER: D

293. The data analysis process steps are ask, prepare, process, analyze, share, and act. During which phase would a data analyst use spreadsheets or query languages to transform data in order to draw conclusions?

A. Act

B. Prepare

C. Process

D. Analyze

ANSWER: D

294. The data analysis process steps are ask, prepare, process, analyze, share, and act. In which step would a data analyst use visuals such as charts or graphs to simplify complex data for better understanding?

A. Process

B. Act

C. Share

D. Prepare

ANSWER: C

295. A data analyst shares insights from their analysis during a formal presentation. In a slideshow, they make a data-driven recommendation for how to solve a business problem. What step of the data analysis process would come next?

A. Process

B. Prepare

C. Ask

D. Act

ANSWER: D

296. Identify among the following which is a correct interpretation of a low standard deviation value for a data distribution?

A. Data is spread over a large range of values

B. Data points are close to the 25th percentile

C. Data points are close to the mode

D. Data points are close to the mean

ANSWER: D

297. Identify among the following which is not true about scatter plots?

A. It is used to identify relationship between attributes

B. It is used to identify clusters

C. It is used to identify outliers

D. It is used in the case of univariate distribution

ANSWER: D

298. Identify among the following that is true if the value of a dissimilarity measure is zero for two objects?

A. The two objects are moderately similar

B. The two objects are very dissimilar

C. The two objects are moderately dissimilar

D. The two objects are very similar

ANSWER: D

299. Identify the type of error that is returned by the following statement?

```
import pandas as pnd
pnd.Series([11,21,31,41], index = ['a','b','c'])
```

A. SyntaxError

B. IndexError

C. ValueError

D. None of the above

ANSWER: C

300. Identify among the following statement that will display values more than 40 from Series ‘S1’?

A. >>>S1[S1 > 40]

B. >>> S1 > 40

C. >>>S1

D. None of the above

ANSWER: A

301. John is working in an IT company. His boss asked him to prepare a chart in python. He is confused that which library will support plotting graph and data visualization in python. Help him to find the library

A. Pandas

B. Math

C. Numpy

D. Matplotlib

ANSWER: D

302. An output of series ‘P1’ is shown below, _____ are the data values in ‘P1 and _____ are the data labels in ‘P1’.

Output:

Feb 2

Mar 3

Apr 4

dtype: int64

A. Month name, Numbers

B. Numbers, Numbers

C. Month name, Month name

D. Numbers, Month name

ANSWER: D

303. How many times value ’10’ will be displayed in the given series ‘P1’?

import pandas as pd

304. P1=pd.Series(10, index = range(1, 10, 3))

print(P1)

A. 1

B. 4

C. 5

D. 3

ANSWER: D

305. David performed addition of series ‘P1’ and ‘P2’ and store the result in series ‘P3’. Both the series ‘P1’ and ‘P2’ have five mismatching index value. How many NaN will be there in ‘P3’?

A. 5

B. 3

C. 4

D. 2

ANSWER: A

306. Identify among the following statement that return Filtered result?

import pandas as pd

S1=pd.Series([11, 12, 5, 6,9])

print(S1) #Statement 1

print(S1>7) #Statement 2

print(S1[S1>7]) #Statement 3

A. Statement 3

B. Statement 2

C. Statement 1

D. None of the above

ANSWER: A

307. Jenny is a class XII student. She is learning "Series" in Python. She is confused in "parameter of Series( ) function". Help her to find the incorrect parameter.

A. number

B. index

C. data

D. dtype

ANSWER: A

308. In given code dataframe 'D1' has _____ rows and _____ columns.

```
import pandas as pd
LoD = [{'a':10, 'b':20}, {'a':5, 'b':10, 'c':20},{'a':7, 'd':10, 'e':20}]
D1 = pd.DataFrame(LoD)
```

A. 3, 5

B. 3, 4

C. 3, 3

D. None of the above

ANSWER: A

309. In given code dataframe 'D1' has _____ rows and _____ columns.

```
import pandas as pd
LoD = {"Name" : ["Amit", "Anil","Ravi"], "RollNo" : [1,2,3]}
D1 = pd.DataFrame(LoD)
```

A. 3, 3

B. None of the above

C. 2, 3

D. 3, 2

ANSWER: D

310. Write the code to remove duplicate row labelled as 'R1' from DataFrame 'DF1'

A. DF1 = DF1.del('R1', axis = 1)

B. DF1 = DF1.drop('R1', axis = 1)

C. DF1 = DF1.del('R1', axis = 0)

D. DF1 = DF1.drop('R1', axis = 0)

ANSWER: D

311. The following statement is _____

```
>>> DF=DF.rename({'Maths':'Sub1','Science':'Sub2'}, axis='index') #DF is a DataFrame
```

A. Error

B. altering the column labels

C. altering the row and column labels (both)

D. altering the row labels

ANSWER: D

312. Evaluate the result of the statement >>>df.shape , if df has the following structure.

	Name	Class	Rollno
0	Amit	6	1
1	Anil	7	2
2	Ravi	8	3

A. (3, 3)

B. (4, 3)

C. (3, 4)

D. None of the above

ANSWER: A

313. Evaluate the result of the statement >>>df.size , if df has the following structure:

	Name	Class	Rollno
0	Amit	6	1
1	Anil	7	2
2	Ravi	8	3

A. None of the above

B. 12

C. 6

D. 9

ANSWER: D

314. Identify the statement to “use * as marker” . (consider plt as an alias name of matplotlib.pyplot)

A. None of the above

B. plt.plot(h, w, mark = “*”) #h and w are data representing x axis and y axis

C. plt.marker(“*”)

D. plt.plot(h, w, marker = “*”) #h and w are data representing x axis and y axis

ANSWER: D

315. To which category Multivariate linear regression belongs?

A. Supervised learning

B. Both supervised and unsupervised learning

C. Neither supervised nor unsupervised learning

D. Unsupervised learning

ANSWER: A

316. The learner is trying to predict housing prices based on the size of each house and number of bedrooms. What type of regression is this?

A. Multivariate Logistic Regression

B. Multivariate Linear Regression

C. Linear Regression

D. Logistic Regression

ANSWER: B

317. Identify the following statements which is false about Ensemble learning?

A. It is an unsupervised learning algorithm

B. More random algorithms can be used to produce a stronger ensemble

C. It is a supervised learning algorithm

D. Ensembles can be shown to have more flexibility in the functions they can represent

ANSWER: A

318. Identify among the following statements which is not true about multi-class classification?

A. An input can belong to one of K classes

B. Each input belongs to more than one class

C. Each training data associated with class labels which is a number from 1 to K

D. Each input belongs to exactly one class

ANSWER: B

319. If machine learning model output involves target variable then that model is called as

A. descriptive model

B. all of the above

C. reinforcement learning

D. predictive model

ANSWER: D

320. If machine learning model output doesnot involve target variable then that model is called as

A. reinforcement learning

B. predictive model

C. descriptive model

D. all of the above

ANSWER: C

321. Imagine a Newly-Born starts to learn walking. It will try to find a suitable policy to learn walking after repeated falling and getting up. Specify what type of machine learning is best suited?

A. classification

B. reinforcement learning

C. kmeans algorithm

D. regression

ANSWER: B

322. Identify among the following methods that we use to find the best fit line for data in Linear Regression?

A. Logarithmic Loss

B. Maximum Likelihood

C. Least Square Error

D. Both A and B

ANSWER: C

323. Identify the common feature selection methods in regression task?

A. all of the above

B. greedy algorithms

C. correlation coefficient

D. none of these

ANSWER: A

324. Regarding bias and variance, which of the following statements are true? (Here 'high' and 'low' are relative to the ideal model.

i. Models which overfit are more likely to have high bias

ii. Models which overfit are more likely to have low bias

iii. Models which overfit are more likely to have high variance

iv. Models which overfit are more likely to have low variance

A. i and ii

B. ii and iii

C. none of these

D. iii and iv

ANSWER: D

325. Identify what feature scaling does before applying K-Mean algorithm?

A. none of these

B. you always get the same clusters. if you use or dont use feature scaling

C. in manhattan distance it is an important step but in euclidian it is not

D. in distance calculation it will give the same weights for all features

ANSWER: D

326. Identify among the following statement(s) what can be true post adding a variable in a linear regression model?

1. R-Squared and Adjusted R-squared both increase

2. R-Squared increases and Adjusted R-squared decreases

3. R-Squared decreases and Adjusted R-squared decreases

4. R-Squared decreases and Adjusted R-squared increases

A. None of the above

B. 1 and 3

C. 2 and 4

D. 1 and 2

ANSWER: D

327. In SVM, RBF kernel with appropriate parameters to perform binary classification where the data is non-linearly separable. In this scenario

A. the decision boundary in the transformed feature space in non-linear

B. the decision boundary in the original feature space in linear

C. the decision boundary in the original feature space in not considered

D. the decision boundary in the transformed feature space in linear

ANSWER: D

328. Identify among the following sentences which are correct in reference to Information gain?

a. It is biased towards single-valued attributes

b. It is biased towards multi-valued attributes

c. ID3 makes use of information gain

d. The approach used by ID3 is greedy

A. a and b

B. a and d

C. all of the above

D. b, c and d

ANSWER: D

329. Even if there are no actual supervisors _____ learning is also based on feedback provided by the environment.

A. Supervised

B. None of the above

C. Unsupervised

D. Reinforcement

ANSWER: D

330. 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 following case?

A. Feature F1 is an example of nominal variable.

B. None of these

C. It doesn't belong to any of the above category.

D. Feature F1 is an example of ordinal variable.

ANSWER: D

331. Say, you have 5000 different models with their predictions and want to ensemble predictions of best x models. Now, which of the following can be a possible method to select the best x models for an ensemble?

A. step wise forward selection

B. step wise backward elimination

C. none of above

D. both

ANSWER: D

332. In terms of bias and variance. Which of the following is true when you fit degree 2 polynomial?

A. bias will be high, variance will be low

B. bias will be low, variance will be high

C. bias will be high, variance will be high

D. bias will be low, variance will be low

ANSWER: A

333. Identify among the following assumptions what do we make while deriving linear regression parameters?

1. The true relationship between dependent y and predictor x is linear

2. The model errors are statistically independent

3. The errors are normally distributed with a 0 mean and constant standard deviation

4. The predictor x is non-stochastic and is measured error-free

A. 1, 2 and 3

B. All of above

C. 1 and 3

D. 1,3 and 4

ANSWER: B

334. A student Grade is a variable F1 which takes a value from A,B,C and D. Which of the following is True in the following case?

A. variable f1 is an example of nominal variable

B. it belongs to both ordinal and nominal categories

C. it doesnt belong to any of the mentioned categories

D. variable f1 is an example of ordinal variable

ANSWER: D

335. Identify the naive assumption in a Naive Bayes Classifier.

A. all the features of a class are conditionally dependent on each other

B. all the features of a class are independent of each other

C. the most probable feature for a class is the most important feature to be cinsidered for classification

D. all the classes are independent of each other

ANSWER: A

336. We usually use feature normalization before using the Gaussian kernel in SVM. What is true about feature normalization?

1. We do feature normalization so that new feature will dominate other

2. Sometimes, feature normalization is not feasible in case of categorical variables

3. Feature normalization always helps when we use Gaussian kernel in SVM

A. 1

B. 2 and 3

C. 1 and 3

D. 1 and 2

ANSWER: D

337. The minimum time complexity for training an SVM is O(n2). According to this fact, what sizes of datasets are not best suited for SVM's?

A. Medium sized datasets

B. Small datasets

C. Large datasets

D. Size does not matter

ANSWER: C

338. Say, you got a situation where you find that your linear regression model is under fitting the data. In such situation which of the following options would you consider?

1. I will add more variables

2. I will start introducing polynomial degree variables

3. I will remove some variables

A. 1 and 3

B. 2 and 3

C. 1 and 2

D. 1, 2 and 3

ANSWER: C

339. Identify the truth about an ensemble classifier?

1. Classifiers that are more "sure" can vote with more conviction

2. Classifiers can be more "sure" about a particular part of the space

3. Most of the times, it performs better than a single classifier

A. All of the above

B. 1 and 3

C. 2 and 3

D. 1 and 2

ANSWER: A

340. Generally, an ensemble method works better, if the individual base models have _____ ?

Note: Suppose each individual base models have accuracy greater than 50%.

A. correlation does not have any impact on ensemble output

B. high correlation among predictions

C. less correlation among predictions

D. None of the above

ANSWER: C

341. Identify the way to counter over-fitting in decision tree?

A. both by pruning the longer rules' and ' by creating new rules'

B. by creating new rules

C. by pruning the longer rules

D. over-fitting is not possible

ANSWER: C

342. Say you are using a Linear SVM classifier with 2 class classification problem. Now you have been given the following data in which some points are circled red that are representing support vectors.If you remove the following any one red points from the data. Does the decision boundary will change?

A. No

B. Yes

C. Can’t say

D. None of the above

ANSWER: B

343. Identify among the following what is true about SVM?

1. Kernel function map low dimensional data to high dimensional space.

2. It is a similarity Function

A. 1 is true, 2 is true

B. 1 is false, 2 is true

C. 1 is true, 2 is false

D. 1 is false, 2 is false

ANSWER: A

344. Identify the statement which is true about prediction problems?

A. the resultant model is designed to classify current behavior.

B. the output attribute must be numeric.

C. the output attribute must be categorical.

D. None of the above

ANSWER: A

345. _____ is much more difficult because it's necessary to determine a supervised strategy to train a model for each feature and, finally, to predict their value

A. removing the whole line

B. All of the above

C. using an automatic strategy to input them according to the other known values

D. creating sub-model to predict those features

ANSWER: D

346. Identify among the following what is an example of feature extraction?

A. construction bag of words from an email

B. forward selection

C. removing stop words

D. applying pca to project high dimensional data

ANSWER: D

347. Identify the steps for using a gradient descent algorithm?

1. Calculate error between the actual value and the predicted value

2. Reiterate until you find the best weights of network

3. Pass an input through the network and get values from output layer

4. Initialize random weight and bias

5. Go to each neurons which contributes to the error and change its respective values to reduce the error

A. 1, 2, 3, 4, 5

B. 5, 4, 3, 2, 1

C. 3, 2, 1, 5, 4

D. 4, 3, 1, 5, 2

ANSWER: D

348. What is the problem in multi regression?

A. both multicollinearity & overfitting

B. overfitting

C. multicollinearity

D. underfitting

ANSWER: A

349. Bayes theorem describes the probability of an event, based on prior knowledge of conditions that might be related to the event.

A. FALSE

B. TRUE

C. Can't say

D. None of the above

ANSWER: B

350. The process becomes a _____ if there is only a discrete number of possible outcomes (called categories)

A. Regression

B. Categories

C. Tree

D. Classification

ANSWER: D

351. Identify among the following metrics which can be used for evaluating regression models?

1. R Squared

2. Adjusted R Squared

3. F Statistics

4. RMSE / MSE / MAE

A. 1, 2, 3 and 4

B. 1 and 2

C. 2, 3 and 4

D. 2 and 4

ANSWER: A

352. During the treatment of cancer patients, the doctor needs to be very careful about which patients need to be given chemotherapy. Which metric should we use in order to decide the patients who should be given chemotherapy?

A. call

B. recall

C. precision

D. score

ANSWER: C

353. The SVM's are less effective when:

A. The data is noisy and contains overlapping points

B. The data is clean and ready to use

C. The data is linearly separable

D. None of these

ANSWER: A

354. The selling price of a house depends on many factors. For example, it depends on the number of bedrooms, number of kitchen, number of bathrooms, the year the house was built, and the square footage of the lot. Given these factors, predicting the selling price of the house is an example of ____ task.

A. binary classification

B. multiple linear regression

C. simple linear regression

D. multilabel classification

ANSWER: B

355. 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 its slope

D. by intercept

ANSWER: C

356. Gaussian distribution when plotted, gives a bell shaped curve which is symmetric about the _____ of the feature values.

A. Variance

B. Mean

C. Discrete

D. Random

ANSWER: B

357. Identify among the following which is the most appropriate kernel that can be used with SVM to separate the classes.

A. linear kernel

B. None of the above

C. polynomial kernel

D. gaussian rbf kernel

ANSWER: D

358. Does the number of features decreases while using feature selection on the data?

A. No

B. Yes

C. Can’t say

D. None of the above

ANSWER: B

359. Generally, which of the following method(s) is used for predicting continuous dependent variable?

1. Linear Regression

2. Logistic Regression

A. 1 and 2

B. None of the above

C. only 2

D. only 1

ANSWER: D

360. If X and Y in a regression model are totally unrelated,

A. the correlation coefficient would be -1

B. None of these

C. the coefficient of determination would be 1

D. the coefficient of determination would be 0

ANSWER: D

361. A company has build a kNN classifier that gets 100% accuracy on training data. When they deployed this model on client side it has been found that the model is not at all accurate. Which of the following thing might gone wrong? Note: Model has successfully deployed and no technical issues are found at client side except the model performance

A. can't say

B. it is probably a underfitted model

C. it is probably a overfitted model

D. wrong client data

ANSWER: C

362. Say, you are working with categorical feature(s) and you have not looked at the distribution of the categorical variable in the test data. You want to apply one hot encoding (OHE) on the categorical feature(s). What challenges you may face if you have applied OHE on a categorical variable of train dataset?

A. All categories of categorical variable are not present in the test dataset.

B. Frequency distribution of categories is different in train as compared to the test dataset.

C. Both A and B

D. Train and Test always have same distribution.

ANSWER: C

363. The cost parameter in the SVM means:

A. The tradeoff between misclassification and simplicity of the model

B. The kernel to be used

C. The number of cross-validations to be made

D. None of the above

ANSWER: A

364. Identify among the following statement which is true about k-NN algorithm?

1. k-NN performs much better if all of the data have the same scale

2. k-NN works well with a small number of input variables (p), but struggles when the number of inputs is very large

3. k-NN makes no assumptions about the functional form of the problem being solved

A. 1,2 and 3

B. 1 and 3

C. only 1

D. 1 and 2

ANSWER: A

365. Simple regression assumes a _____ relationship between the input attribute and output attribute.

A. linear

B. inverse

C. quadratic

D. reciprocal

ANSWER: A

366. How does number of observations influence overfitting?

1. In case of fewer observations, it is easy to overfit the data.

2. In case of fewer observations, it is hard to overfit the data.

3. In case of more observations, it is easy to overfit the data.

4. In case of more observations, it is hard to overfit the data.

A. None of these

B. 2 and 3

C. 1 and 3

D. 1 and 4

ANSWER: D

367. Bayes Theorem is given by where

1. P(H) is the probability of hypothesis H being true.

2. P(E) is the probability of the evidence(regardless of the hypothesis).

3. P(E|H) is the probability of the evidence given that hypothesis is true.

4. P(H|E) is the probability of the hypothesis given that the evidence is there.

A. Can't say

B. FALSE

C. TRUE

D. None of the above

ANSWER: C

368. Say you are using RBF kernel in SVM with high Gamma value. What does this signify?

A. The model would consider even far away points from hyperplane for modeling

B. None of the above

C. The model would not be affected by distance of points from hyperplane for modeling

D. The model would consider only the points close to the hyperplane for modeling

ANSWER: D

369. What does dimensionality reduction reduce?

A. stochastics

B. performance

C. collinearity

D. entropy

ANSWER: C

370. Identify the role of FacetGrid in Seaborn.

A. Enables users to create a grid of subplots based on the values of one or more categorical variables.

B. Used to create multiple plots

C. None of the above

D. Both A and B

ANSWER: D

371. Identify the use cases of PairGrid?

A. Exploratory data analysis

B. Visualizing high-dimensional data

C. All of the above

D. Comparing multiple subgroups

ANSWER: C

372. Identify the significance of hue parameters in Seaborn plots.

A. Used to assign colors to different categories or groups in the data.

B. Helpful for distinguishing between various subgroups and identifying patterns or trends in the data.

C. None of the above

D. Both A and B

ANSWER: D

373. The Central Limit Theorem says that the sampling distribution of the sample mean is approximately normal if _____

A. all possible samples are selected.

B. None of the above.

C. the standard error of the sampling distribution is small.

D. the sample size is large.

ANSWER: D

374. The Central Limit Theorem says that the mean of the sampling distribution of the sample means is ___

A. exactly equal to the population mean.

B. close to the population mean if the sample size is large.

C. equal to the population mean divided by the square root of the sample size.

D. None of the above

ANSWER: A

375. The Central Limit Theorem says that the standard deviation of the sampling distribution of the sample means is _____

A. None of the above

B. close to the population standard deviation if the sample size is large.

C. exactly equal to the standard deviation.

D. equal to the population standard deviation divided by the square root of the sample size.

ANSWER: D

376. Samples of size 25 are selected from a population with mean 40 and standard deviation 7.5. The mean of the sampling distribution of sample means is _____ according to central limit theorem.

A. 40

B. 8

C. 7.5

D. None of the above

ANSWER: A

377. Samples of size 25 are selected from a population with mean 40 and standard deviation 7.5. The standard error of the sampling distribution of sample means is _____ according to central limit theorem.

A. 0.3

B. None of the above

C. 7.5

D. 1.5

ANSWER: D

378. Identify which of the following is not important in determining data quality?

A. Database

B. Consistency

C. Completeness

D. Accuracy

ANSWER: A

379. Identify which of the following is not a form of data transformation?

A. Compression

B. Discretization

C. Concept hierarchy

D. Normalization

ANSWER: A

380. What is data mining?

A. Deleting unnecessary data

B. Extracting useful patterns or information from large datasets

C. Storing data securely

D. Sorting data alphabetically

ANSWER: B

381. Identify which of the following is not a basic data mining task?

A. Classification

B. Prediction

C. Spooling

D. Clustering

ANSWER: C

382. Identify which of the following is not an issue in data mining?

A. High dimensionality

B. Outliers

C. Overfitting

D. Shortage of data

ANSWER: D

383. Identify which of the following is a subset of data warehouse focused on a specific functional area?

A. Flat files

B. Association rules

C. Data mart

D. Database

ANSWER: C

384. An object refers to the feature of the data.

A. False

B. True

C. Can't say

D. None of the above

Answer: A

385. Identify which among the following are not among Various Operations in Data Warehousing?

A. Roll up

B. Dice

C. Drill down

D. Sticking

ANSWER: D

386. Identify which of the following refers to the set of features that describe a data object?

A. Sample

B. Instance

C. Attribute vector

D. Data point

ANSWER: C

387. By which parameters are the data sets made up of?

A. Data Relations

B. Data Patterns

C. Data Classes

D. Data Objects

ANSWER: D

388. If the value to the variable is 55 degrees, into which type of attribute can the data be classified?

A. Numeric Attribute

B. Binary Attribute

C. Ordinal Attribute

D. Nominal Attribute

ANSWER: A

389. Choose the basic building block of a data set

A. Data mining

B. Association rules

C. Cluster

D. Data object

ANSWER: D

390. For a data object stored in a database, the attributes of the data object are represented by _____

A. Relation schema

B. Rows of the database

C. Columns of the database

D. Relational index

ANSWER: C

391. Identify which of the following does not describe the properties of a data object?

A. Attributes

B. Features

C. Dimensions

D. Instances

ANSWER: D

392. Identify the measure of central tendency that separates the data set into higher and lower halves is _____

A. Mean

B. Midrange

C. Mode

D. Median

ANSWER: D

393. NumPy provides a function _____ similar to range that returns arrays instead of lists, to create sequences of numbers.

A. aline

B. aspace

C. arange

D. all of the mentioned

ANSWER: C

394. Choose the correct statement.

A. NumPy main object is the homogeneous multidimensional array

B. In Numpy, dimensions are called axes

C. All of the mentioned

D. Numpy array class is called ndarray

ANSWER: C

395. Identify which of the following is used to calculate the mean of a NumPy array?

A. min()

B. max()

C. mean()

D. All of the above

ANSWER: C

396. Identify which of the following is used to reshape a NumPy array?

A. None of the options

B. argmax()

C. Both A and B

D. reshape()

ANSWER: D

397. Identify which of the following are modules/libraries in Python?

A. NumPy

B. All of the above

C. Matplotlib

D. Pandas

ANSWER: B

398. Identify an important library used for analyzing data.

A. Pandas

B. Random

C. Math

D. None of the above

ANSWER: A

399. Crucial data structure of pandas is/are _____

A. Series

B. Data Frame

C. None of the above

D. Both of the above

ANSWER: D

400. Identify which of the following library in Python is used for plotting graphs and visualization.

A. Matplotlib

B. NumPy

C. Pandas

D. None of the above

ANSWER: A

401. Identify which of the following command is used to install pandas?

A. pip pandas

B. install pandas

C. pip install pandas

D. None of the above

ANSWER: C

402. Which from the options is a one-dimensional array.

A. Data Frame

B. None of the above

C. Both of the above

D. Series

ANSWER: D

403. Starting from what a Series by default have numeric data labels

A. 0

B. 2

C. 1

D. 3

ANSWER: A

404. Identify the index value of "Alice" in the following Series?

```
import pandas as pd
S1 = pd.Series["John", "Sam", "Alice", "Bob", "Bill"]
```

A. 0

B. 1

C. 2

D. 3

ANSWER: C

405. From _____ we can create DataFrame

A. Numpy arrays

B. All of the above

C. Dictionary of Lists

D. List of Dictionaries

ANSWER: B

406. Identify which of the following is used to display first 2 rows of DataFrame 'DF'?

A. DF.head(2)

B. DF.header(2)

C. DF.head( )

D. None of the above

ANSWER: A

407. Identify which of the following function is used to load the data from the CSV file into a DataFrame?

A. read.csv( )

B. read_csv( )

C. csvread( )

D. csvr( )

ANSWER: B

408. Identify the problem when sometimes the data under analysis has too many attributes, some of which may not be useful as per the task being performed.

A. Overfitting

B. Missing data

C. Outliers in data

D. Dimensionality curse

ANSWER: D

409. A decision tree is a flowchart like structure and has nodes and branches. What does branches and nodes represent, respectively?

A. Attribute value and outcome of the test

B. Class Description and Classes

C. Classes and Class Description

D. Outcome of the test and attribute value

ANSWER: D

410. Identify which values are generally modeled by Regression Analysis?

A. Distorted Class Labels

B. Discrete Class Labels

C. Unordered Class Labels

D. Continuous Valued Functions

ANSWER: D

411. Identify which of the following refers to the distribution of data based on two variables?

A. Time series

B. Univariate distribution

C. Bivariate distribution

D. Association rules

ANSWER: C

412. Identify which of the following attribute values have a meaningful order but no information about the magnitude between successive values?

A. Ordinal

B. Binary

C. Nominal

D. Numeric

ANSWER: A

413. What is it called when the original data can be reconstructed from the compressed data without any information loss.

A. Lost data transformation

B. Lossy data reduction

C. Loss-full data transformation

D. Lossless data reduction

ANSWER: D

414. The _____ function returns its argument with a modified shape, whereas the _____ method modifies the array itself.

A. reshape2, resize3

B. resize, reshape

C. reshape, resize

D. all of the mentioned

ANSWER: C

415. Evaluate the result of the following code?

```
import numpy as np
a = np.arange(10)
print(a[1:4])
```

A. [1, 2, 3]

B. [0, 1, 2]

C. [5, 6, 7]

D. [2, 4, 6]

ANSWER: A

416. Evaluate the result of the following code?

```
import numpy as np
a = np.array([4, 5, 6])
b = np.array([1, 2, 3])
c = np.vstack((a, b))
print(c)
```

A. [[4, 5, 6],[1, 2, 3]]

B. [[1, 4], [2, 5], [3, 6]]

C. [1, 2, 3, 4, 5, 6]

D. Error

ANSWER: A

417. Evaluate the result of the following code?

```
import numpy as np
a = np.array([7, 8, 9])
b = np.array([4, 5, 6])
c = np.concatenate((a, b))
print(c)
```

A. [[1, 2, 3], [4, 5, 6]]

B. [[1, 4], [2, 5], [3, 6]]

C. [7, 8, 9, 4, 5, 6]

D. Error

ANSWER: C

418. Identify which of the following statement is wrong?

A. We can create Series from Dictionary in Python.

B. Keys of dictionary become index of the series.

C. Order of indexes created from Keys may not be in the same order as typed in dictionary.

D. Series is an interface

ANSWER: D

419. If given code is not returning any error, then how many values will be there in arr?

```
# import pandas as pd
>>> sr1 = pd.Series(arr, index = ["Sep", "Oct", "Nov", "Dec"])
```

A. 1

B. 2

C. 3

D. 4

ANSWER: D

420. In Series we can access elements by using _____ index and _____index.

A. Positional, labelled

B. Positional, Naming

C. Numeric, labelled

D. None of the above

ANSWER: A

421. Identify which of the following statement will print Series 'srdemo1' in reverse order?

A. print(srdemo1[: : 1])

B. print(srdemo1[: : -1])

C. print(srdemo1[-1: : 1])

D. None of the above

ANSWER: B

422. Identify which of the following statement will display the difference of two Series 'SR1' and 'SR2'?

A. >>>SR1 – SR2

B. >>>SR1.sub(SR2)

C. Both of the above

D. None of the above

ANSWER: C

423. What 'data' in the following code could be?

S1 = pd.Series(data)

A. All of the above

B. Scalar value

C. Python dictionary

D. Python sequence

ANSWER: A

424. Choose the correct statement :

Statement1 : A Numpy array requires homogeneous data.

Statement2 : Pandas DataFrame can have heterogeneous data.

A. Both the statements are correct

B. Statement2 is correct

C. Statement1 is correct

D. Both the statements are wrong

ANSWER: A

425. The following code create a dataframe named 'T1' with how many columns?

import pandas as pd

LS1 = [{'x':10, 'y':20}, {'x':5, 'y':10, 'z':20}]

T1 = pd.DataFrame(LS1)

A. 1

B. 2

C. 3

D. 4

ANSWER: C

426. In following code dataframe 'T1' has _____ rows and _____ columns.

import pandas as pd

S1 = pd.Series([21, 22, 23, 24], index = ['a', 'b','c','d'])

S2 = pd.Series([11, 23, 33, 44], index = ['a', 'bb','c','dd'])

T1 = pd.DataFrame([S1,S2])

A. 2, 4

B. 4, 6

C. 4, 4

D. 2, 6

ANSWER: D

427. T1[:] = 31 , will set _____ values of a Data Frame 'T1' to 31.

A. Only First Row

B. Only First Column

C. None of the above

D. All

ANSWER: D

428. What the following statement will perform?

df = df.drop(['N1', 'N2', 'N3'], axis = 1) #df is a DataFrame object

A. delete three columns having labels 'N1', 'N2' and 'N3'

B. delete three rows having labels 'N1', 'N2' and 'N3'

C. delete any three columns

D. return error

ANSWER: A

429. DataFrame.loc[ ] is an important method that is used for _____ with DataFrames

A. Both of the above

B. Boolean based indexing

C. Label based indexing

D. None of the above

ANSWER: C

430. Transpose the DataFrame means _____

A. Both of the above

B. Doubling the number of rows in DataFrame

C. Row indices and column labels of the DataFrame replace each other's position

D. None of the above

ANSWER: C

431. Name of the figure is passed to what function as parameter.

A. savefig( )

B. show( )

C. plot( )

D. None of the above

ANSWER: A

432. Identify which of the following pyplot function is used to set the label for the x-axis.

A. xlabeled( )

B. None of the above

C. x_axis_label( )

D. xlabel( )

ANSWER: D

433. To show the grid lines in plot, we can write _____ #plt is an alias of matplotlib.pyplot

A. Both of the above

B. plt.grid(True)

C. plt.grid( )

D. None of the above

ANSWER: A

434. In matplotlib, 'marker' is an attribute of what function?

A. show( )

B. None of the above

C. display( )

D. plot( )

ANSWER: D

435. The mango learning algorithm is based on a dataset that consists of three variables – color, softness, tastiness of the mango. Which is more likely to be the target variable?

A. mango

B. Softness

C. Tastiness

D. Color

ANSWER: C

436. What happens due to overfitting?

A. Hypothesis works well on training data but works poorly on test data

B. Hypothesis works well on training data and works well on test data

C. Hypothesis works poorly on training data but works well on test data

D. Hypothesis works poorly on training data and works poorly on test data

ANSWER: A

437. scikit-learn offers the class _____ , which is responsible for filling the holes using a strategy based on the mean, median, or frequency

A. LabelEncoder

B. Imputer

C. DictVectorizer

D. LabelBinarizer

ANSWER: B

438. Covariance between two attributes is termed as variance when the two attributes are _____

A. Binary

B. Different

C. Identical

D. Nominal

ANSWER: C

439. Evaluate the result of the following code?

```
import numpy as np
a = np.array([[5, 6], [7, 8]])
print(a.ndim)
```

A. 0

B. 1

C. 2

D. 3

ANSWER: C

440. Evaluate the result of the following code?

```
import numpy as np
a = np.array([7, 8, 9])
b = np.array([1, 2, 3])
c = a + b
print(c)
```

A. [1, 2, 3, 4, 5, 6]

B. [[1, 4], [2, 5], [3, 6]]

C. [8, 10, 12]

D. Error

ANSWER: C

441. Identify which of the following is used to find the indices of the maximum and minimum elements in a NumPy array?

A. amax() and amin()

B. max() and min()

C. argmax() and argmin()

D. None of the above

ANSWER: C

442. Evaluate the result of the following code?

```
import numpy as np
a = np.array([[1, 2], [3, 4]])
b = np.array([[5, 6], [7, 8]])
c = np.dot(a, b)
print(c)
```

A. Error

B. [[5, 6], [7, 8], [1, 2], [3, 4]]

C. [[1, 5], [2, 6], [3, 7], [4, 8]]

D. [[19, 22], [43, 50]]

ANSWER: D

443. Identify what type of error is returned by following code?

```
import pandas as pd
S1 = pd.Series(data = (11, 2, -2), index = [6, 8, 2, 1])
print(S1)
```

A. SyntaxError

B. IndexError

C. ValueError

D. None of the above

ANSWER: C

444. Identify which of the following statement return Boolean result?

```
import pandas as pd
P1=pd.Series([11, 12, 5, 6,9])
print(P1) #Statement 1
print(P1>6) #Statement 2
print(P1[P1>7]) #Statement 3
```

A. Statement 1

B. Statement 2

C. Statement 3

D. None of the above

ANSWER: B

445. John is working in an IT firm as Data Manager. He stored the salary of all employees in Series named "emp". His Boss asked him to filter the employees whose salary is more than 20000. Help him to find the correct code. Sample of data stored is shown below.

0 25000

1 20000

2 21000

3 30000

A. print(emp > 20000)

B. print(emp (emp> 20000))

C. print(emp [emp > 20000])

D. print[emp> 20000]

ANSWER: C

446. Maggie wants to create a series named 'P1'. He has written the following codes. Her friend Amy checked the code and said that one of the code given below is not working. As a friend of Maggie, help her to find the incorrect code.

A. P1=pd.Series(data=[11, 12, 5, 6,9],index=[1,2,3,4,5])

B. P1=pd.Series([11, 12, 5, 6,9],index=[1,2,3,4,5])

C. P1=pd.Series([11, 12, 5, 6,9],[1,2,3,4,5])

D. P1=pd.Series(data=[11, 12, 5, 6,9], [1,2,3,4,5])

ANSWER: D

447. Billy has written few points about iloc( ) function of Series in Python. His friend Peter told that one of the written statement is not correct. Help him to find the incorrect statement.

A. In iloc( ) method, we have to pass an integer index.

B. S1.iloc[:3] will display first three values of Series 'S1'

C. S1.iloc[3] will display fourth value of Series 'S1'

D. This method include the last element of the range passed.

ANSWER: D

448. Mr. Walker is working in an IT company. He stored the salaries of all the employees of January month in Series 'Jan_Sal' and salaries of February month in Series 'Feb_Sal'. Now he wants to add the salaries of both months. He has written the following statement. Identify the correct one.

A. print(Feb_Sal plus Jan_Sal)

B. print(Feb_Sal_add_Jan_Sal)

C. print(Feb_Sal + Jan_Sal)

D. None of the above

ANSWER: C

449. Write a statement to delete a column having column label as 195 in dataframe DF.

A. print(DF.drop(195,axis=0))

B. None of the above

C. All of the above

D. print(DF.drop(195,axis=1))

ANSWER: D

450. The learner is trying to predict the cost of apple based on its size. The variable "cost" is _____

A. independent variable

B. categorical variable

C. ranked variable

D. target Variable

ANSWER: D

451. Identify the goal of gradient descent?

A. Reduce complexity

B. Minimize cost function

C. Maximize cost function

D. Reduce overfitting

ANSWER: B

452. Identify what happens when the learning rate is high?

A. Most of the times, it overshoots the minima

B. It overshoots the maxima

C. It always reaches the minima quickly

D. Nothing happens

ANSWER: A

453. Identify what happens when the learning rate is low?

A. It always reaches the minima quickly

B. Nothing happens

C. It overshoots the minima

D. It reaches the minima very slowly

ANSWER: D

454. Identify the factor on which the updating of each parameter dependent on?

A. The learning rate and the target variable

B. Target variable

C. The number of training examples

D. The learning rate

ANSWER: A

455. Identify what is updated by gradient descent after each iteration?

A. The learning rate

B. The number of training examples

C. Target variable

D. Independent variables

ANSWER: D

456. Which among the following adopts a dictionary-oriented approach, associating to each category label a progressive integer number.

A. DictVectorizer

B. LabelBinarizer class

C. LabelEncoder class

D. FeatureHasher

ANSWER: C

457. Identify the False option. You trained a binary classifier model which gives very high accuracy on the training data, but much lower accuracy on validation data.

A. this is an instance of overfitting

B. the training and testing examples are sampled from different distributions

C. the training was not well regularized

D. this is an instance of under fitting

ANSWER: D

458. What is it called when the data label associated with a particular value of Series?

A. Data value

B. None of the above

C. Value

D. Index

ANSWER: D

459. Identify which among the following library is to be imported to create Series.

A. NumPy

B. None of the above

C. Matplotlib

D. Pandas

ANSWER: D

460. Identify which among the following function/method help to create Series?

A. seriesCreate( )

B. None of the above

C. createSeries( )

D. Series( )

ANSWER: D

461. The left most column show _____ value when you print/display any series.

A. Value

B. Data

C. Index

D. None of the above

ANSWER: C

462. Pandas Series is analogous as a _____ in a spreadsheet

A. Table

B. Cell

C. Column

D. None of the above

ANSWER: C

463. Identify among the following statement which will modify the first three values of Series 'S1'?

A. S1[0, 1, 2] = 100

B. All of the above

C. S1[: 3] = 100

D. S1[0 : 3] = 100

ANSWER: B

464. Identify the index value of 31 in given Series 'S1'?

```
import pandas as pd
S1=pd.Series([1,2,31,4], index = ['a','b','c','d'])
```

A. 'c'

B. 2

C. error

D. Both A and B

ANSWER: D

465. Python libraries contain a collection of built-in _____

A. Data

B. Data Structure

C. Packages

D. Modules

ANSWER: D

466. Identify what is "Pandas" in the following statement?

```
import pandas as pd
```

A. Package

B. Module

C. Library

D. Function

ANSWER: C

467. Identify among the following library which help to visualize data?

A. Matplotlib

B. Numpy

C. Pandas

D. Random

ANSWER: A

468. Identify among the following attribute of Series which is used to set the name of Series object?

A. size

B. values

C. index.name

D. name

ANSWER: D

469. Identify among the following attribute which is used to check NaN value in Series?

A. none

B. HasNans

C. Hasnans

D. hasnans

ANSWER: D

470. head( ) function return _____ n rows and tail function return _____ n rows from a pandas object.

A. last, first

B. first, last

C. last, seven

D. first, second

ANSWER: B

471. What function is used to sort a Series object on the basis of values?

A. sort.values( )

B. sort_Values

C. none

D. sort_values( )

ANSWER: D

472. Identify the statement that will assign a name to the Series 'S1'.

A. >>> S1[name] = "Empl"

B. >>> S1_name = "Empl"

C. >>> S1.name = "Empl"

D. >>> S1.indexname = "Empl"

ANSWER: C

473. plot() function in matplotlib plots a _____ by default.

A. Line chart

B. Bar graph

C. Histogram

D. Pie chart

ANSWER: A

474. Identify among the following pyplot function which is used to set a title for the chart.

A. title( )

B. c_title( )

C. setTitle( )

D. None of the above

ANSWER: A

475. Identify the attribute of plot( ) function which help to specify the type of chart?

A. type

B. types

C. kindof

D. kind

ANSWER: D

476. Identify the statement to plot a line chart for data stored in a DataFrame 'df'.

A. df.plot(kind = "line")

B. plot(df, line)

C. df.plot(line)

D. None of the above

ANSWER: A

477. Identify the attribute of plot( ) function which is used to set the width of line in line plot in matplotlib?

A. widthline

B. none of the above

C. widthofline

D. linewidth

ANSWER: D