

179 min left

6. C2 linear Regression

[Previous](#)[Next](#)

Course 2

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

What does standardised scaling do?

[Clear Ans](#)

Answer Options

Select any one option

- Bring all data points in the range 0 to 1
- Bring all data points in the range -1 to 1
- Bring all the data points in a normal distribution with mean 0 and standard deviation 1
- Bring all the data points in a normal distribution with mean 1 and standard deviation 0

<<

37

Course 3

1 2 3

180 min left

5. C2 Logistic Regression

[Previous](#)[Next](#)

Course 2

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

Which of the following is true for weight of evidence (WoE) analysis?

Answer Options

Select any one option

[Clear Ans](#)

It helps in finding the different predictive patterns for the different segments that might be present in the data.

WoE helps in treating missing values for both continuous and categorical variables.

WoE values should follow an increasing or decreasing trend across bins.

All of the above

Course 3

1 2 3

180 min left

4. C2 Logistic Regression

[Previous](#)[Next](#)

Course 2

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

<<

28 29 30

31 32 33

34 35 36

37

If you use a random number generator to predict the output 0 or 1 for a binary classification problem, what will be the area under the curve of the ROC curve?

Answer Options

Select any one option

[Clear Ans](#) 0 0.5 1 100

3. C2 linear Regression

Course 2

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

<<

28 29 30

31 32 33

34 35 36

37

Course 3

1 2 3

Consider the following two assumptions for a simple regression model. (Assume X and y to be independent and dependent variables respectively).

Statement 1: There is a linear relationship between X and y.

Statement 2: X and y are normally distributed.

Answer Options

Select any one option

[Clear Ans](#) Statement 1 is correct and Statement 2 is incorrect Statement 1 is incorrect and Statement 2 is correct Both the statements are correct Both the statements are incorrect

180 min left

2. C2 Multiple correct answer

[Previous](#)[Next](#)

Course 2

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

Answer Options

Select one or more options

[Clear Ans](#) Dendrogram inspection method Elbow Method Single Linkage Method Silhouette score

Course 3

1 2 3

36. C2 Multiple correct answer

Course 2

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

<<

28 29 30

31 32 33

34 35 36

36

37

Course 3

1 2 3

ROC curve shows the tradeoff between the True Positive Rate (TPR) and the False Positive Rate (FPR). TPR and FPR are sensitivity and specificity respectively. The following function is written in Python using metrics package from the scikit-learn library for a ROC curve function.

```
def draw_roc(actual, probs):
    fpr tpr,thresholds=metrics.roc_curve(actual,probs,drop_intermediate=False)
    auc_score = metrics.roc_auc_score(actual, probs)
    return None
```

Which of the following statements are true? (More than one option may be correct.)

Answer Options

Select one or more options

Clear Ans

 The area under the ROC curve can be more than 1. The arguments passed in the above function are actual values of the target variable and the predicted values (i.e., 0 or 1) Larger the area under the curve, the better will be the model The arguments passed in the above function are actual values of the target variable and the respective predicted probabilities

35. C2 Clustering

Course 2

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

<<

28 29 30

31 32 33

34 35 36

37

Course 3

1 2 3

Silhouette metric for any i th point is given by: $S(i) = (b(i) - a(i)) / \max\{b(i), a(i)\}$

Which of the following is not true about Silhouette metric?

Answer Options

Select any one option

Clear Ans

 b(i) is the average distance from the nearest neighbour cluster(Separation). a(i) is the average distance from own cluster(Cohesion). If $S(i) = 1$ then the datapoint is similar to its own cluster. Silhouette Metric ranges from 0 to +1

34. C2-Basics of NLP and Text Mining

[Previous](#)[Next](#)

Course 2

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

Choose the correct option from the following.

The difference between '+' and '*' quantifier is ____.

Answer Options

Select any one option

[Clear Ans](#) '+' needs the preceding character to be present at least once whereas '*' does not need the same. '*' needs the character to be present at least once whereas '+' does not need the same. Both the quantifiers have the same functionality. None of the above

Course 3

1 2 3

177 min left

33. C2-Decision Trees

[Previous](#)[Next](#)

Course 2

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8
- 9
- 10
- 11
- 12
- 13
- 14
- 15
- 16
- 17
- 18
- 19
- 20
- 21
- 22
- 23
- 24
- 25
- 26
- 27
- 28
- 29
- 30
- 31
- 32
- 33
- 34
- 35
- 36
- 37

Suppose you train a decision tree with the following data. Which feature should we split on at the root?

Answer Options

Select any one option

[Clear Ans](#) X Y Z Cannot be determined

Course 3

- 1
- 2
- 3

177 min left

32. C2-Basics of NLP and Text Mining

[Previous](#)[Next](#)

Course 2

- [1](#)
- [2](#)
- [3](#)
- [4](#)
- [5](#)
- [6](#)
- [7](#)
- [8](#)
- [9](#)
- [10](#)
- [11](#)
- [12](#)
- [13](#)
- [14](#)
- [15](#)
- [16](#)
- [17](#)
- [18](#)
- [19](#)
- [20](#)
- [21](#)
- [22](#)
- [23](#)
- [24](#)
- [25](#)
- [26](#)
- [27](#)
- [««](#)
- [32](#)
- [33](#)
- [34](#)
- [35](#)
- [36](#)
- [37](#)

Course 3

- [1](#)
- [2](#)
- [3](#)

Which of the following strings will match with the regular expression "^01*0\$"?

- 1. 0
- 2. 00
- 3. 0111111110

Answer Options

Select any one option

[Clear Ans](#)

- Only option 1
- Only option 3
- Both 1&2
- Both 2&3

177 min left

31. C2 linear Regression

[Previous](#)[Next](#)

Course 2

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

31 32 33

34 35 36

37

A scatterplot was plotted for two variables - age and income to find out how the income depends on the age of a person. It was found that as the income increases linearly with age, the variability in income also increases. This is a violation of which of the following assumptions of linear regression?

Answer Options

Select any one option

[Clear All](#) Homogeneity Heterogeneity Homoskedasticity Linearity

Course 3

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

180 min left

1. C2 Business Problem Solving

[Previous](#)[Next](#)

Course 2

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27



Select any one option

[Clear Ans](#)

28 29 30

 3>1>5>2>4>7>6>8

31 32 33

 3>2>1>5>4>7>6>8

34 35 36

 4>3>1>2>5>7>6>8

37

 3>2>1>5>4>7>8>6

Course 3

1 2 3

177 min left

30. C2 Logistic Regression

[Previous](#)[Next](#)

Course 2

- [1](#)
- [2](#)
- [3](#)
- [4](#)
- [5](#)
- [6](#)
- [7](#)
- [8](#)
- [9](#)
- [10](#)
- [11](#)
- [12](#)
- [13](#)
- [14](#)
- [15](#)
- [16](#)
- [17](#)
- [18](#)
- [19](#)
- [20](#)
- [21](#)

Recall the telecom churn example, if the log odds for churn are equal to 0 for a customer, then that means -

- [22](#)
- [23](#)
- [24](#)
- [25](#)
- [26](#)
- [27](#)
- [28](#)
- [29](#)
- [30](#)
- [31](#)
- [32](#)
- [33](#)
- [34](#)
- [35](#)
- [36](#)
- [37](#)

Answer Options

Select any one option

[Clear All](#)

- There is no chance of the customer churning
- The probability of the customer churning is equal to the probability of the customer not churning
- The probability of the customer churning is very small compared to the probability of the customer not churning
- The probability of the customer churning is very large compared to the probability of the customer not churning

Course 3

- [1](#)
- [2](#)
- [3](#)
- [4](#)
- [5](#)
- [6](#)
- [7](#)
- [8](#)
- [9](#)
- [10](#)
- [11](#)
- [12](#)
- [13](#)
- [14](#)
- [15](#)
- [16](#)
- [17](#)
- [18](#)
- [19](#)
- [20](#)
- [21](#)
- [22](#)
- [23](#)
- [24](#)

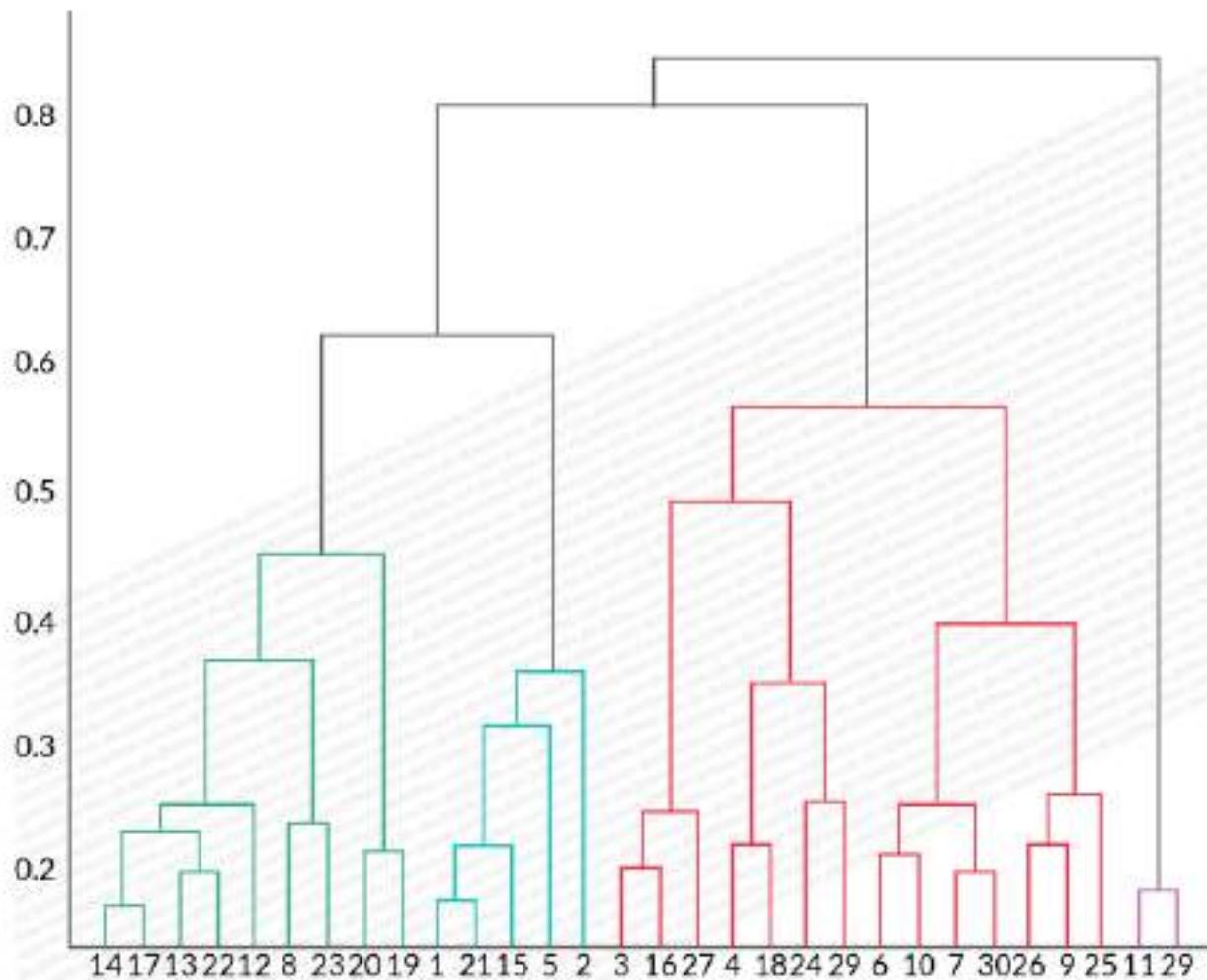
Dataset 2

1	1	3
4	5	6
7	8	9
10	11	12
13	14	15
16	17	18
19	20	21
22	23	24
25	26	27
28	29	30
31	32	33
34	35	36
37		

Dataset 1

1	2	3
4	5	6
7	8	9
10	11	12
13	14	15
16	17	18
19	20	21
22	23	24
25	26	27
28	29	30
31	32	33
34	35	36
37		

You obtained the following dendrogram after performing K-means clustering on a dataset. Which of the following statements can be concluded from the dendrogram?



Answer Options

Select any one option

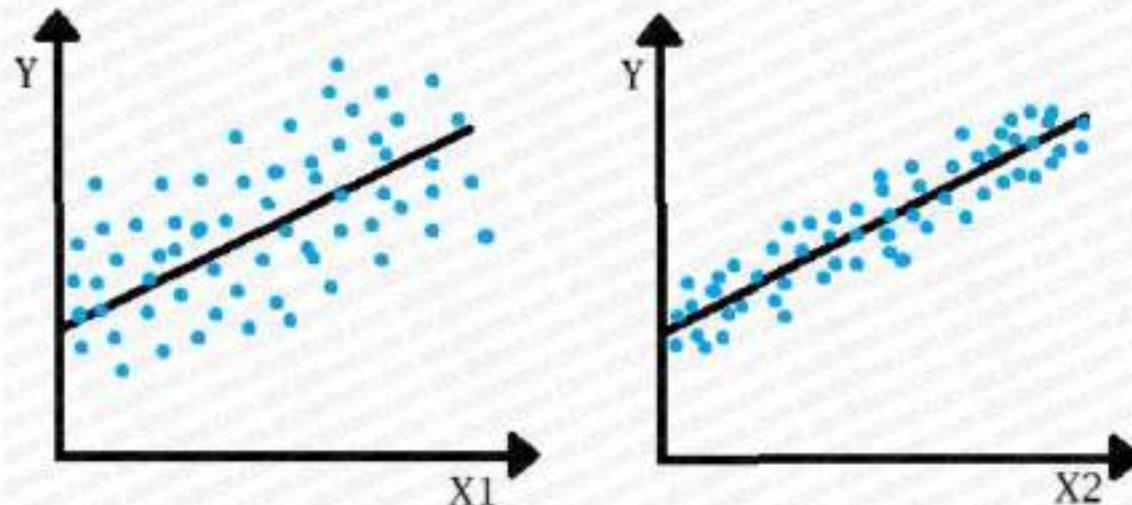
- The initial number of clusters is 6
- There are 25 data points used in the above clustering algorithm
- Single linkage is used to define the distance between two clusters in the above dendrogram

28. C2 linear Regression

Course 2

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8
- 9
- 10
- 11
- 12
- 13
- 14
- 15
- 16
- 17
- 18
- 19
- 20
- 21
- 22
- 23
- 24
- 25
- 26
- 27
- 28
- 29
- 30
- 31
- 32
- 33
- 34
- 35
- 36

For the same dependent variable Y, two models were created using the independent variables X1 and X2. The following two graphs represent the fitted line on the scatterplot. (Both of the graphs are on the same scale)



Which of the following is true about the residuals in these two models?

Answer Options

Select any one option

Clear A

Course 3

The sum of residuals in model 2 is higher than model 1

The sum of residuals in model 1 is higher than model 2

?

Both have the same sum of residuals

Nothing can be said about the sum of residuals from the graph

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8
- 9
- 10
- 11
- 12
- 13
- 14
- 15
- 16
- 17
- 18

178 min left

28. C2 linear Regression

[Previous](#)[Next](#)

Course 2

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30



31 32 33

34 35 36

37

Course 3

1 2 3

4 5 6

7 8 9

For the same dependent variable Y, two models were created using the independent variables X1 and X2. The following two graphs represent the fitted on the scatterplot. (Both of the graphs are on the same scale)

Which of the following is true about the residuals in these two models?

Answer Options

Select any one option

[Clear All](#) The sum of residuals in model 2 is higher than model 1 The sum of residuals in model 1 is higher than model 2 Both have the same sum of residuals Nothing can be said about the sum of residuals from the graph

178 min left

27. C2-Basics of NLP and Text Mining

< Previous

Next

Course 2

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30



31 32 33

34 35 36

37

Answer Options

Select any one option

Clear Ans

3

4

5

6

Course 3

1 2 3

4 5 6

7 8 9

10 11 12

178 min left

26. C2 Multiple correct answer

[Previous](#)[Next](#)

Course 2

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

Answer Options

Select one or more options

[Clear All](#) from sklearn.linear_model import LogisticRegression

lr = LogisticRegression()

lr.fit(X_train, y_train)

 import statsmodel.api as sm

lr = sm.GLM(y_train,(sm.add_constant(X_train)),

family = sm.families.Binomial())

lr.fit()

Course 3

1 2 3

4 5 6

7 8 9

 from sklearn.linear_model import LogisticRegression

lr = LogisticRegression()

lr.predict(X_train, y_train)

 import statsmodel.api as sm

178 min left

25. C2-Decision Trees

[Previous](#)[Next](#)

Course 2

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8
- 9
- 10
- 11
- 12
- 13
- 14
- 15
- 16
- 17
- 18
- 19
- 20
- 21
- 22
- 23
- 24
- 25
- 26
- 27
- 28
- 29
- 30
- 31
- 32
- 33
- 34
- 35
- 36
- 37

Which of the following metrics measures how often a randomly chosen element would be incorrectly identified?

Answer Options

Select any one option

[Clear All](#) Entropy Information Gain Gini Index

Course 3

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8
- 9

 None of these

178 min left

24. C2 linear Regression

[Previous](#)[Next](#)

Course 2

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30



31 32 33

34 35 36

37

Course 3

1 2 3

4 5 6

7 8 9

Which of the following is true regarding the error terms in linear regression?

Answer Options

Select any one option

[Clear All](#) The sum of residuals should be zero The sum of residuals should be lesser than zero The sum of residuals should be greater than zero There is no such restriction on what the sum of residuals should be

178 min left

23. C2 Logistic Regression

[Previous](#)[Next](#)

Course 2

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

31 32 33

34 35 36

37

Given an imbalanced dataset, the ratio of positive to negative class is 1:10000. You run a logistic regression model and find out that the model has a high value of precision and a low value of recall. Which of the following statements is true?

Answer Options

Select any one option

[Clear All](#) The class is handled well by the data The model is not able to detect the class, but when it does it is highly trustable The model is able to detect the class but it includes data points from the other class as well

1 2 3

4 5 6

7 8 9

10 11 12

Course 3

 The class is handled poorly by the data

178 min left

22. C2 Logistic Regression

[Previous](#)[Next](#)

Course 2

- 1 2 3
- 4 5 6
- 7 8 9
- 10 11 12
- 13 14 15
- 16 17 18
- 19 20 21
- 22** 23 24
- 25 26 27
- 28 29 30
- 31 32 33
- 34 35 36
- 37

Consider the following confusion matrix.

Total=500	Actual Positive	Actual Negative
Predicted Positive	196	20
Predicted Negative	28	256

Which among the following is the lowest for the given confusion matrix?

Answer Options

Select any one option

[Clear Ans](#) Accuracy Precision Sensitivity Specificity

Course 3

- 1 2 3
- 4 5 6
- 7 8 9
- 10 11 12

178 min left

21. C2 linear Regression

[Previous](#)[Next](#)

Course 2

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30



31 32 33

34 35 36

37

How is regression different from classification?

Answer Options

Select any one option

[Clear All](#) One is supervised while the other is unsupervised One is iterative while the other is closed form In regression, the response variable is numeric while it is categorical in classification

Course 3

1 2 3

4 5 6

7 8 9

 None of the above

178 min left

20. C2 Clustering

[Previous](#)[Next](#)

Course 2

Which of the following statements is NOT true?

- 1 2 3
- 4 5 6
- 7 8 9
- 10 11 12
- 13 14 15
- 16 17 18

Answer Options

- 19 20 21
- 22 23 24
- 25 26 27
- 28 29 30
- 31 32 33
- 34 35 36
- 37

Select any one option

[Clear Ans](#)

- Each time the clusters are made during the K-means algorithm, the centroid is updated.
- The cluster centres that are computed in the K-means algorithm are given by centroid value of the cluster points.
- Standardization of the data is not important before applying Euclidean distance as a measure of similarity/dissimilarity.

Course 3

- 1 2 3
- 4 5 6
- 7 8 9
- 10 11 12

- The centroid of a column with data points 25, 32, 34 and 23 is 28.5

176 min left

19. C2 Multiple correct answer

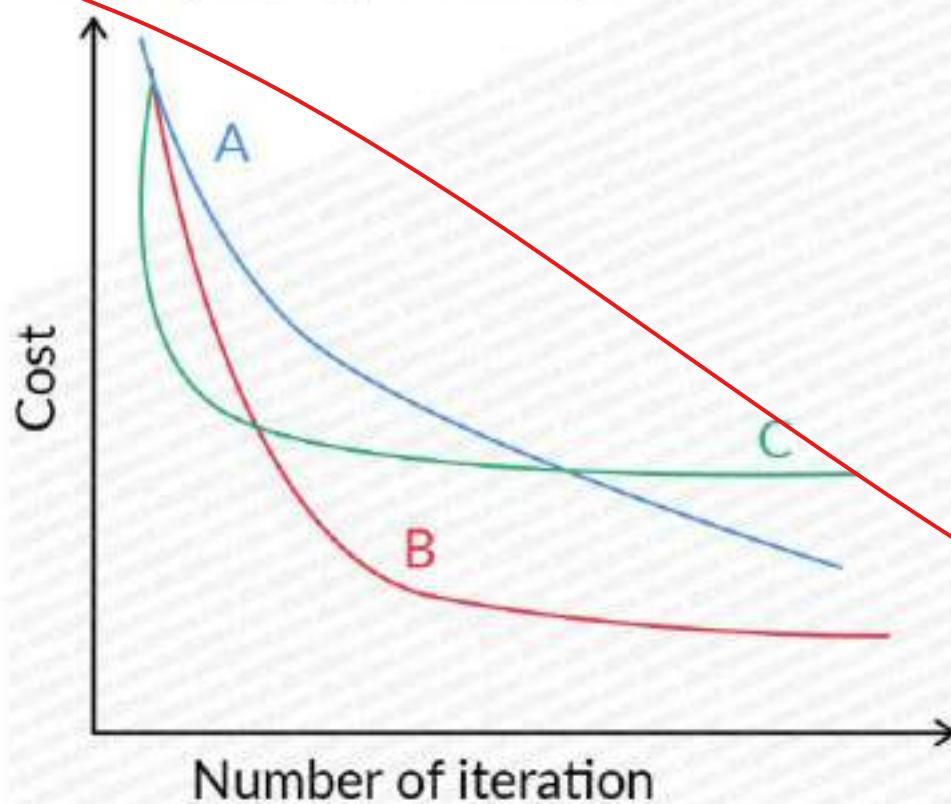
< Previous

Course 2

- | | | |
|----|----|----|
| 1 | 2 | 3 |
| 4 | 5 | 6 |
| 7 | 8 | 9 |
| 10 | 11 | 12 |
| 13 | 14 | 15 |
| 16 | 17 | 18 |
| 19 | 20 | 21 |
| 22 | 23 | 24 |
| 25 | 26 | 27 |
| 28 | 29 | 30 |
| 31 | 32 | 33 |
| 34 | 35 | 36 |
| 37 | | |

Course 3

Observe the following cost function graph with different learning rates.



Which of the following statements are true about the learning rate? (More than one option may be correct.)

Answer Options:

Select one or more options.

Clear

 The learning rate of curve C is highest among all curves The learning rate for curve B is lower than A The learning rate for curve B is higher than A

179 min left

19. C2 Multiple correct answer

[Previous](#)[Next](#)

Course 2

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

<<

28 29 30

31 32 33

34 35 36

37

Course 3

1 2 3

Observe the following cost function graph with different learning rates.

Which of the following statements are true about the learning rate? (More than one option may be correct.)

Answer Options

Select one or more options

[Clear Ans](#) The learning rate of curve C is highest among all curves The learning rate for curve B is lower than A The learning rate for curve B is higher than A The learning rate of curve C is the smallest among all curves None of the above

179 min left

18. C2 Logistic Regression

[Previous](#)[Next](#)

Course 2

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27



28 29 30

31 32 33

34 35 36

37

Course 3

1 2 3

You have built a Logistic Regression model that is trying to predict whether a loan is approved or not based on a person's FICO score. Here are the model parameters: Intercept (B_0) = -9.346 and coefficient of FICO score = 0.0146. Given these parameters, can you calculate the probability of a loan getting approved for someone with a FICO score of 655?

Answer Options

Select any one option

[Clear Ans](#) 0.35 0.45 0.55 0.65

179 min left

17. C2 linear Regression

[Previous](#)[Next](#)

Course 2

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27



28 29 30

31 32 33

34 35 36

37

Course 3

1 2 3

Suppose you run a regression with one of the feature variables T, with all the remaining feature variables. The R-squared of this model was found out to be 0.8. What will be the VIF for the variable T?

Answer Options

Select any one option

[Clear Ans](#) 1.56 2.77 3.33 5.00

16. C2 Multiple correct answer

[Previous](#)[Next](#)

Course 2

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

Answer Options

Select one or more options

[Clear Ans](#) The null hypothesis for a simple linear regression model is $H_0: \beta_1 = 0$ If the p-value turns out to be greater than 0.05 for β_1 , it means β_1 is significant If β_1 turns out to be insignificant, that means there is no relationship between the dependent and independent variable

37

Course 3

1 2 3

 If the p-value turns out to be less than 0.05 for β_0 , it means that β_0 is non-zero

179 min left

15. C2 linear Regression

[Previous](#)[Next](#)

Course 2

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

In linear regression, the metric F-statistic is used to determine

Answer Options

Select any one option

[Clear Ans](#) the significance of the individual beta coefficients the variance explanation strength of the model the significance of the overall model fit Both A & C

37

Course 3

1 2 3

179 min left

14. C2 Logistic Regression

[Previous](#)[Next](#)

Course 2

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

Which of the following is correct for a logistic regression model?

Answer Options

Select any one option

[Clear Ans](#) The independent variables should not be multicollinear. The dependent variable should follow Normal Distribution. The log odds in a logistic regression model lies between 0 and 1. F1-score is always the best metric for evaluating a logistic regression model.

Course 3

1 2 3

13. C2 Business Problem Solving

Course 2

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

<<

 Statement 1 is correct and Statement 2 is wrong

28 29 30

31 32 33

34 35 36

37

Course 3

1 2 3

 Both the statements are correct None of the statements are correct

The coronavirus disease (COVID-19), was declared a pandemic by the World Health Organization (WHO) in February 2020. Currently, there are no vaccines or treatments that have been officially approved by WHO after clinical trials. India has not seen the peak of infection yet and the number of infections is touching a new height daily. The business unit of an Indian health and hygiene company approaches you to know "**Why the sales of masks is decreasing despite the number of corona infections increasing daily**".

Answer the below question:

Consider the following two statements:

Statement 1: Understanding the change in customer behaviour is an important factor to be considered for business understanding for the problem statement above

Statement 2: One of the possible hypotheses for the above problem statement: There is a rise in the number of companies manufacturing normal/surgical masks due to which the sales of the client's company is decreasing.

[Clear Ans](#)

179 min left

12. C2 Clustering

[Previous](#)[Next](#)

Course 2

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

Which of the following is not true for Hopkins statistic?

Answer Options

Select any one option

[Clear Ans](#) Hopkins statistic decides if the data is suitable for clustering or not Hopkins statistic lie between -1 and 1 If the Hopkins statistic comes out to be 0, then the data is uniformly distributed If the Hopkins statistic comes out to be 1, then the data highly suitable for clustering

Course 3

1 2 3

179 min left

11. C2 Logistic Regression

[Previous](#)[Next](#)

Course 2

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

<<

28 29 30

31 32 33

34 35 36

37

Course 3

1 2 3

Consider the following univariate logistic model:

$$Y = \beta_0 + \beta_1 X_1$$

Which of the following statements is NOT true?

Answer Options

Select any one option

[Clear Ans](#) The maximum likelihood estimation determines the best combination of β_0 and β_1 . If β_1 is increased by 1 unit, Y increases by 1 unit. β_0 is the y-intercept If β_1 is increased by 1 unit, log-odds increases by 1 unit.

10. C2 Business Problem Solving

Course 2

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27



Logistic regression

28 29 30

31 32 33

34 35 36

37

Course 3

1 2 3

The coronavirus disease (COVID-19), was declared a pandemic by the World Health Organization (WHO) in February 2020. Currently, there are no vaccines or treatments that have been officially approved by WHO after clinical trials. India has not seen the peak of infection yet and the number of infections is touching a new height daily. The business unit of an Indian health and hygiene company approaches you and know "Why the sales of masks is decreasing despite the number of corona infections increasing daily". Answer the following questions:

Suppose you mapped the above problem statement with a classification problem, either a customer will buy a mask or not. Your will build _____ model as your initial solution.

Answer Options

Select any one option

[Clear Ans](#) Neural Network Logistic regression Decision tree All of the above

179 min left

9. C2 Clustering

[Previous](#)[Next](#)

Course 2

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27



28 29 30

31 32 33

34 35 36

37

Initialising the following command in Python will result in the following: `model_clus = KMeans(n_clusters = 6, max_iter=50)`

Answer Options

Select any one option

[Clear Ans](#) Run maximum 6 iterations Run maximum 40 iterations Create 6 final clusters Create 50 final clusters

Course 3

1 2 3

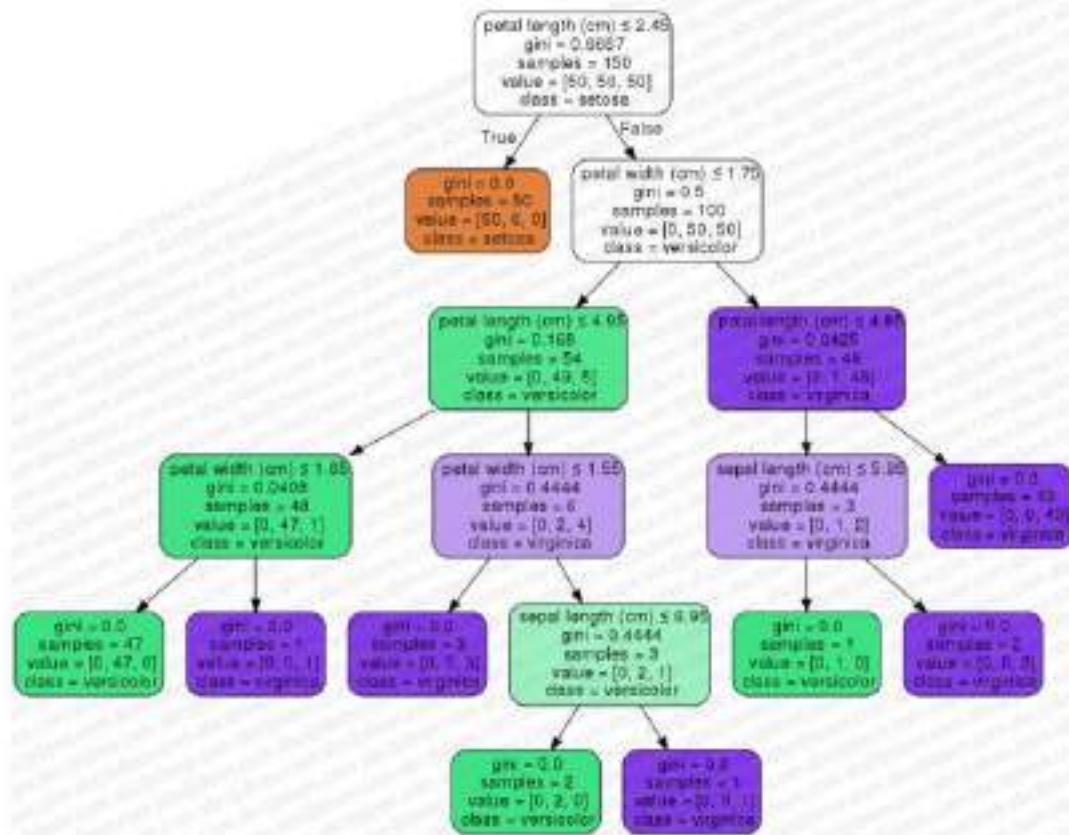
Course 2

- 1 2 3
 4 5 6
 7 8 9
 10 11 12
 13 14 15
 16 17 18
 19 20 21
 22 23 24
 25 26 27
 28 29 30
 31 32 33
 34 35 36

Course 3

- 1 2 3
 4 5 6
 7 8 9
 10 11 12
 13 14 15
 16 17 18
 19 20 21
 22 23 24
 25 26 27
 28 29 30
 31 32 33

Refer to the decision tree given below and choose the statement that is correct as per this tree.



Answer Options

Select any one option

Clear

The tree given above will show very good performance on the train data.

The tree given above is an underfitting tree.

If the petal length is more than 2.45, then it is equally likely that the flower is either setosa or virginica.

179 min left

8. C2-Decision Trees

[Previous](#)[Next](#)

Course 2

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8
- 9
- 10
- 11
- 12
- 13
- 14
- 15
- 16
- 17
- 18
- 19
- 20
- 21
- 22
- 23
- 24
- 25
- 26
- 27
- 28
- 29
- 30
- 31
- 32
- 33
- 34
- 35
- 36
- 37

Refer to the decision tree given below and choose the statement that is correct as per this tree.

Answer Options

Select any one option

[Clear Ans](#)

- The tree given above will show very good performance on the train data.
- The tree given above is an underfitting tree.
- If the petal length is more than 2.45, then it is equally likely that the flower is either setosa or virginica.
- Both B and C

Course 3

- 1
- 2
- 3

179 min left

7. GroupBy and OrderBy

[Previous](#)[Next](#)

SQLITE

Course 2

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

31 32 33

34 35 36

37

Course 3

1 2 3

INPUT TABLE

You're given a **orders** table and the columns in the orders table are shown below:

orders	
Order_Id	INT
Type	VARCHAR(10)
Real_Shipping_Days	INT
Scheduled_Shipping_Days	INT
Customer_Id	INT
Order_City	VARCHAR(20)
Order_Date	DATE
Order_Region	VARCHAR(15)
Order_State	VARCHAR(20)
Order_Status	VARCHAR(20)
Shipping_Mode	VARCHAR(20)
Indexes	

QUERY

- Calculate count of all the orders
 - **where** the *Order_State* is **Gujarat**
 - **where** the *Order_Status* is **PENDING**.
 - **Note** - Use the alias of **oc** for count of orders.
- **Group the results by** *Order_City*
- **Order them by** *oc & Order_City* in **ascending order**.

OUTPUT COLUMNS

oc, Order_City

Console

Custom Test Case

178 min left

21. C2 Logistic Regression

[Previous](#)[Next](#)

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

31 32 33

34 35 36

37 38 39

40 41 42

43 44 45

The output of a logistic model is:

Answer Options

Select any one option

[Clear Ans](#) 0 or 1 Any value between 0 and 1 0.5 Depends on the business problem

178 min left

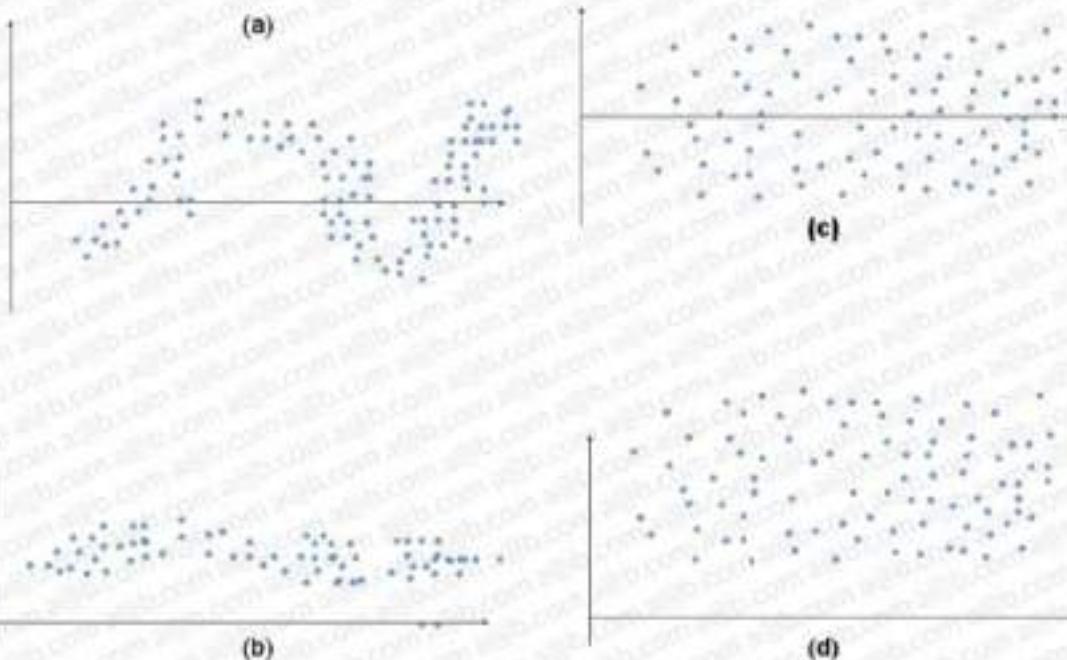
20. C2 linear Regression

[Previous](#)[Next](#)

MCQs

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8
- 9
- 10
- 11
- 12
- 13
- 14
- 15
- 16
- 17
- 18
- 19
- 20
- 21
- 22
- 23
- 24
- 25
- 26
- 27
- 28
- 29
- 30
- 31
- 32
- 33
- 34
- 35
- 36
- 37
- 38
- 39
- 40
- 41
- 42
- 43
- 44
- 45

The distribution of error terms in a linear regression model should look like (the horizontal line represents $y=0$):



Answer Options

Select any one option

[Clear Ans](#)

- A
- B
- C

178 min left

18. C2 Logistic Regression

[Previous](#)[Next](#)

MCQs

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8
- 9
- 10
- 11
- 12
- 13
- 14
- 15
- 16
- 17
- 18
- 19
- 20
- 21
- 22
- 23
- 24
- 25
- 26
- 27
- 28
- 29
- 30
- 31
- 32
- 33
- 34
- 35
- 36
- 37
- 38
- 39
- 40
- 41
- 42
- 43
- 44
- 45

Which of the following statements is NOT true?

Answer Options

Select any one option

[Clear Ans](#)

- In the case of a fair coin, the odds of getting heads is 1
- The error values of linear and logistic regression have to be normally distributed
- Specificity decreases with an increase in sensitivity
- As TPR increases, FPR also increases

178 min left

16. C2 linear Regression

[Previous](#)[Next](#)

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

 A, C, and D

28 29 30

31 32 33

34 35 36

37 38 39

40 41 42

43 44 45

Which of the following assumptions do we make while building a simple linear regression model? (assume X and y to be independent and dependent variables respectively).

- A) There is a linear relationship between X and y.
- B) X and y are normally distributed.
- C) Error terms are independent of each other.
- D) Error terms have constant variance.

Answer Options

Select any one option

[Clear Ans](#) A, B, C and D A, C, and D A, B and C B, C and D

178 min left

15. C2 Clustering

[Previous](#)[Next](#)

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

<<

28 29 30

31 32 33

34 35 36

37 38 39

40 41 42

43 44 45

Silhouette metric for any i th point is given by: $S(i) = (b(i) - a(i)) / \max\{b(i), a(i)\}$

Which of the following is not true about Silhouette metric?

Answer Options

Select any one option

[Clear Ans](#) b(i) is the average distance from the nearest neighbour cluster(Separation). a(i) is the average distance from own cluster(Cohesion). If $S(i) = 1$ then the datapoint is similar to its own cluster. Silhouette Metric ranges from 0 to +1

178 min left

12. C2 Logistic Regression

< Previous

Next

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30



31 32 33

34 35 36

37 38 39

40 41 42

43 44 45

46 47 48

49 50 51

52 53 54

Consider the following two statements:

Statement 1: Suppose the value of Precision and Recall for a model are 0.65 and 0.75 respectively. Then the value of F1-score will be ~0.696.**Statement 2:** Mean squared error is a metric that can be used to evaluate a logistic regression model.

Answer Options

Select any one option

Clear Ans

 Statement 1 is wrong and statement 2 is correct Statement 1 is correct and statement 2 is wrong Both the statements are correct None of the statements are correct

179 min left

11. C2 Clustering

[Previous](#)[Next](#)

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

Which of the following statements is NOT true?

19 20 21

Select any one option

[Clear Ans](#)

22 23 24

 The cluster centers that are computed in the K-means algorithm are given by the centroid value of the cluster points.

25 26 27

 Standardization of the data is important before applying Euclidean distance as a measure of similarity/dissimilarity

28 29 30



31 32 33

 The centroid of a column with data points 25, 32, 34 and 23 is 28.5

37 38 39

 The Euclidean distance between two points (10,2) and (4,5) is 7.

40 41 42

43 44 45

46 47 48

49 50 51

52 53 54

179 min left

6. C2 Clustering

[Previous](#)[Next](#)

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30



31 32 33

34 35 36

37 38 39

40 41 42

43 44 45

46 47 48

49 50 51

52 53 54

For a K-means clustering process, the Hopkins statistic for the dataset came out to be 0.8. Hence the dataset is

Answer Options

Select any one option

[Clear Ans](#) Suitable for clustering Not suitable for clustering Can't say from the given information None of the above

179 min left

5. C2 Business Problem Solving

[Previous](#)[Next](#)

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

Any business problem solving will have following steps.(M)

1. To Identify the right data sources, that will be useful in formulating the final solution
2. Develop hypothesis and assess the overall impact of the hypothesized solution
3. Asking the right question for business and problem understanding
4. Define the solution approach: What will be the POC model? What will be the metrics for the model evaluation? etc.
5. Converting the business problem to a data science problem
6. Start your model building process with the simple POC model. And then increase the complexity of the POC model and optimize the parameters to get the best result.
7. Performing EDA on the datasets
8. Model Evaluation.

What will be the correct flow for solving the above/any business problem?

Answer Options

Select any one option

Clear Ans

 3>1>5>2>4>7>6>8 3>2>1>5>4>7>6>8 4>3>1>2>5>7>6>8 3>2>1>5>4>7>8>6

46 47 48

49 50 51

52 53 54

172 min left

67. C2 Logistic Regression

[◀ Previous](#)

MCQs

- 1 2 3
- 4 5 6
- 7 8 9
- 10 11 12
- 13 14 15
- 16 17 18
- 19 20 21

22 23 24 Answer Options

Select any one option

[Clear](#)

- It helps in capturing the seasonal fluctuations that might be present in the data
- It helps to find the optimal cutoff point more easily
- It helps in finding the different predictive patterns for the different set of data points that might be present in the data
- It helps capture the trends easily when there is a class imbalance

- 67 68 69
- 70

Coding

172 min left

65. C2 Multiple correct answer

< Previous

MCQs

- 1 2 3
4 5 6
7 8 9
10 11 12
13 14 15
16 17 18
19 20 21

The ROC curve shows the tradeoff between the True Positive Rate (TPR) and the False Positive Rate (FPR). The following function is written in Python using the metrics package from the scikit-learn library. The code defines a function `draw_roc` that takes `actual` and `probs` as arguments.

```
def draw_roc(actual, probs):
    fpr, tpr, thresholds=metrics.roc_curve(actual,probs,drop_intermediate=False)
    auc_score = metrics.roc_auc_score(actual,probs)
    return None
```

Which of the following statements are true? (More than one option may be correct.)

Answer Options

Select one or more options.

Clear

The arguments passed in the above function are actual values of the target variable and the predicted values (i.e., 0 or 1)

The arguments passed in the above function are actual values of the target variable and the respective predicted probabilities

The area under the ROC can take any value between 0 and 1

Larger the area under the curve, the better will be the model

- 52 53 54
55 56 57
58 59 60
61 62 63

- 64 65 66
67 68 69

122 min left

64. C2 Multiple correct answer

[Previous](#)

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

26 28 27

28 29 30

31 32 33

34 35 36

37 38 39

41 42

43 44 45

46 47 48

50 51

53 54

56 58 57

58 59

61 62 63

64 65 66

67 68 69

70

Coding

Which of the following metrics can be used for finding the appropriate number of clusters in K-means clustering? (More than one option may be correct)

Answer Options

Select one or more options

[Clear](#) Silhouette Score Elbow Curve Hopkins Statistic Dendogram

172 min left

63. C2 Multiple correct answer

[Previous](#)

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

26 28 27

28 29 30

31 32 33

34 35 36

37 38 39

41 42

43 44 45

46 47 48

50 51

53 54

56 58 57

58 59

61 62

64 65 66

67 68 69

70

Which of the following statements are correct in the context of logistic regression? (More than one option may be correct.)

Answer Options

Select one or more options

[Clear](#) The dummies for continuous variables make the model more unstable Weight of evidence (WoE) helps in treating missing values for both continuous and categorical variables WoE should follow a non-monotonic trend across bins. Data clumping can be a problem with transforming continuous variables to dummies. Information value or IV is an important indicator of predictive power.

122 min left

62. C2 Logistic Regression

< Previous

Next >

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

31 32 33

34 35 36

37 38 39

40 41 42

43 44 45

46 47 48

49 50 51

52 53 54

55 56 57

58 59 60

61 62 63

64 65 66

67 68 69

70

Take a look at the following three problem statements.

problem statement 1: Let's say that you are building a telecom churn prediction model with the business objective that your company wants to implement an aggressive customer retention campaign to retain 'high churn-risk' customers. This is because a competitor has launched extremely low-cost mobile plans, and you want to avoid churn as much as possible by incentivising the customers. Assume that budget is not a constraint.

problem statement 2: Let's say you are building a cancer detection model with the objective that both the patient who has cancer and the patient who has not cancer can be detected correctly. It can have serious implications if you predict either of the class wrong. i.e., If wrongly detected as 'not cancer', the patient will die of cancer, and if wrongly detected as 'cancer', the patient will die of chemotherapy.

Problem statement 3: You have to build an image classification model where 60% of images belong to one class and rest 40% images belong to another class. You have to predict the class of a new image.

Which is the correctly matched model evaluation metric for the above classification models?

Answer Options

Select any one option

Clear

 Problem statement 1: Specificity Problem statement 2: Sensitivity Problem statement 2: Specificity Problem statement 3: Accuracy

Coding

172 min left

61. C2 linear Regression

< Previous

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

31 32 33

34 35 36

37 38 39

40 41 42

43 44 45

46 47 48

49 50 51

52 53 54

55 56 57

58 59 60

61 62 63

64 65 66

67 68 69

70

Coding

RFE method is used for:

Answer Options

Select any one option

Clear

 Dummy variable creation Detecting multicollinearity Feature selection Univariate regression

1 2 3

174 min left

60. C2 linear Regression

< Previous

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

Answer Options

25 26 27

Select any one option

Clear

 Mean of residuals of old model > Mean of residuals of new model. Mean of residuals of old model < Mean of residuals of new model. Mean of residuals of old model = Mean of residuals of new model. Information provided is not enough to comment on the mean of residuals.

59 60 61

62 63

64 65 66

67 68 69

70

60

174 min left

59. C2 Multiple correct answer

< Previous

MCQs

- 1 2 3
- 4 5 6
- 7 8 9
- 10 11 12
- 13 14 15
- 16 17 18
- 19 20 21
- 22 23 24

Which of the following statements are true? (More than one option may be correct.)

Answer Options

Select one or more options

Clear

TSS (Total Sum of Squares) is defined as the sum of all squared differences between the observed dependent variable and its mean.

R-squared can take any value between 0 and 1.

Larger the R-squared value, the better the regression model fits the observations.

If RSS = 5.50 and TSS = 11, the value of VIF will be 1.33.



MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30



31 32 33

34 35 36

37 38 39

40 41 42

43 44 45

46 47 48

49 50 51

52 53 54

The coronavirus disease (COVID-19), was declared a pandemic by the World Health Organization (WHO) in February 2020. Currently, there are no vaccines or treatments that have been officially approved by WHO after clinical trials. India has not seen the peak of infection yet and the number of infections is touching a new height daily. The business unit of an Indian health and hygiene company approaches you to know "**Why the sales of masks is decreasing despite the number of corona infections increasing daily**".

Answer the below question:

Consider the following two statements:

Statement 1: Understanding the change in customer behaviour is an important factor to be considered for business understanding for the problem statement above.

Statement 2: One of the possible hypotheses for the above problem statement: There is a rise in the number of companies manufacturing normal/surgical masks due to which the sales of the client's company is decreasing.

Answer Options

Select any one option

Clear Ans

Statement 1 is correct and Statement 2 is wrong

Statement 2 is correct and Statement 1 is wrong

Both the statements are correct

None of the statements are correct

174 min left

52. C2 linear Regression

[Previous](#)[Next](#)

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

Answer Options

Select any one option

[Clear All](#) The r-squared of model-2 will be less than that of model-1 The r-squared of model-2 increases, but the complexity of model-2 also increases The r-squared of model-2 decreases, but the complexity of model-2 also increases Nothing can be said about the r-squared of model-2

49 50 51

52 53 54

51. C2 Clustering

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30



31 32 33

34 35 36

37 38 39

✓ Both the statements are correct

40 41 42

43 44 45

46 47 48

49 50 51

51

52 53 54

Consider the following two Statements:

Statement 1: The distance between 2 clusters is the maximum distance between any 2 points in the clusters in complete linkage.Statement 2: Most of the time Complete linkage will produce unstructured dendograms.

Answer Options

Select any one option

Clear Ans

 Statement 1 is correct and Statement 2 is wrong Statement 1 is wrong and Statement 2 is correct ✓ Both the statements are correct Both the statements are wrong

175 min left

47. C2 linear Regression

[Previous](#)[Next](#)

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

Answer Options

Select any one option

[Clear Ans](#)

- The VIF has a lower bound of 0
- The VIF has no upper bound
- VIF for a variable generally changes if you drop one of the predictor variables
- If a variable is a product of two other variables, it can have a high VIF



46 47 48

49 50 51

52 53 54

175 min left

46. C2 Logistic Regression

[Previous](#)[Next](#)

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

Answer Options

Select any one option

[Clear All](#) 0.35 0.40 0.45 0.50

40 41 42

43 44 45

46 47 48

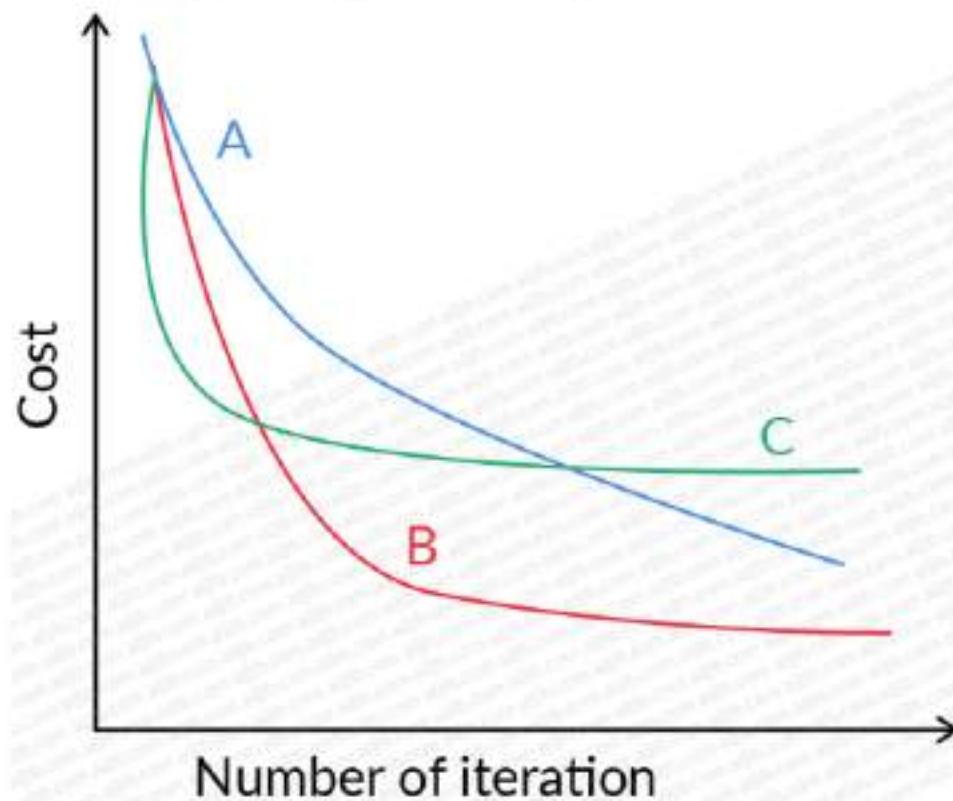
49 50 51

52 53 54

MCQs

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8
- 9
- 10
- 11
- 12
- 13
- 14
- 15
- 16
- 17
- 18
- 19
- 20
- 21
- 22
- 23
- 24
- 25
- 26
- 27
- 28
- 29
- 30
- 31
- 32
- 33
- 34
- 35
- 36
- 37
- 38
- 39
- 40
- 41
- 42

Observe the following cost function graph with different learning rates.



Which of the following statements are true about the learning rate? (More than one option may be correct.)

Answer Options

Select one or more options

Clear

The learning rate of curve C is highest among all curves

The learning rate for curve B is lower than A

The learning rate for curve B is higher than A

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30



31 32 33

34 35 36

37 38 39

40 41 42

43 44 45

46 47 48

49 50 51

52 53 54

Consider the following confusion matrix.

Total=500	Actual Positive	Actual Negative
Predicted Positive	196	20
Predicted Negative	28	256

Which among the following is the highest for the given confusion matrix?

Answer Options

Select any one option

Clear Ans

 Accuracy Precision Sensitivity Specificity

175 min left

43. C2 Logistic Regression

[Previous](#)[Next](#)

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

Recall the telecom churn example. If the log odds for churn are equal to $1/3$ for a customer, then that means -

19 20 21

Answer Options

Select any one option

[Clear Ans](#) The probability of the customer not churning is 3 times the probability of the customer churning The probability of the customer churning is 3 times more than the probability of the customer not churning The probability of the customer not churning is 4 times the probability of the customer churning The probability of the customer churning is 4 times more than the probability of the customer not churning

43 44 45

46 47 48

49 50 51

52 53 54

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

Answer Options

Select any one option

Clear Ans

- Multicollinearity is a problem when your only goal is to predict the independent variable from a set of dependent variables
- Multicollinearity is a problem when your goal is to infer the effect on the dependent variable due to independent variables
- Multicollinearity is not a problem if a variable is not collinear with your variable of interest
- Multicollinearity is not a problem if there are multiple dummy (binary) variables that represent a categorical variable with three or more categories



39

40 41 42

43 44 45

46 47 48

49 50 51

52 53 54

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30



31 32 33

34 35 36

37 38 39

40 41 42

43 44 45

46 47 48

49 50 51

52 53 54

A client approached you with a problem statement. You decided to build a multiple linear regression model on the dataset provided. The dataset consists of 40 features. Obviously all features will not be significant. Selecting the relevant features manually will be a tougher task. You can use RFE to select relevant features. RFE is an automated feature selection technique. Initially, you assumed 25 features can explain your whole data.

Which of the following commands correctly calls the RFE technique in Python? (Here "lm" is the fitted instance of multiple linear regression model)

Answer Options

Select any one option

Clear Ans

- from statsmodel.feature_selection import RFE
rfe = RFE(lm, 25)
rfe = rfe.fit(X_train, y_train)
- from sklearn.feature_selection import RFE
rfe = RFE(lm, 25)
rfe = rfe.predict(X_train, y_train)
- from sklearn.feature_selection import RFE
rfe = RFE(lm, 25)
rfe = rfe.fit(X_train, y_train)
- from RFE import feature_selection
rfe = RFE(lm, 25)
rfe = rfe.predict(X_train, y_train)

176 min left

34. C2 Clustering

[Previous](#)[Next](#)

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

Clustering is used to identify the below?

Answer Options

Select any one option

[Clear All](#) Data distribution Correlation among the data points Principal components Subgroups in the data

34 35 36

37 38 39

40 41 42

43 44 45

46 47 48

49 50 51

52 53 54

176 min left

33. C2 Clustering

[Previous](#)[Next](#)

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

31 32 33

34 35 36

37 38 39

40 41 42

43 44 45

46 47 48

49 50 51

52 53 54

In hierarchical clustering, the shortest distance and the maximum distance between points in two clusters are defined as _____ and _____ respectively.

Answer Options

Select any one option

[Clear Ans](#) Single linkage and Complete linkage Complete linkage and Single linkage Single linkage and Average linkage Complete linkage and Average linkage

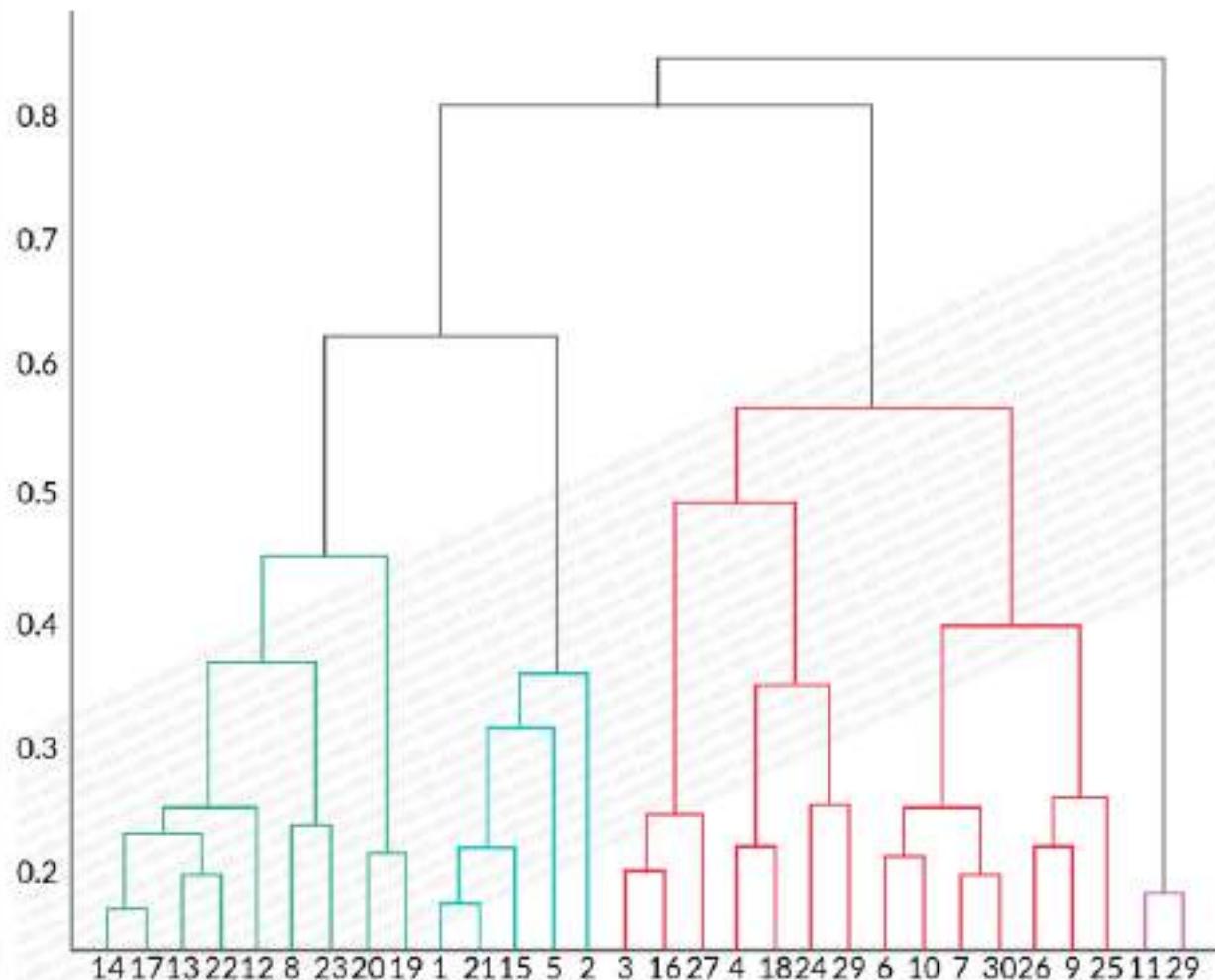
Score:

1	1	3
4	5	6
7	8	9
10	11	12
13	14	10
16	17	18
19	20	21
22	23	24
25	26	27
28	29	30
31	32	
34	35	36
37	38	
39	40	
41	42	
43	44	
46	47	
48	49	
50	51	
53	54	
55	56	
58	59	
60	61	
63	64	
65	66	
67	68	
69	70	

Coding:

1	2	3
4		
5		

You obtained the following dendrogram after performing K-means clustering on a dataset. Which of the following statements can be concluded from the dendrogram?



Answer Options:

Select any one option:

- The initial number of clusters is 6
- There are 25 data points used in the above clustering algorithm
- Single linkage is used to define the distance between two clusters in the above dendrogram
- The elbow dendrogram in tecplot.com is not suitable for K-Means clustering

176 min left

31. C2 linear Regression

[Previous](#)[Next](#)

MCQs:

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

If the coefficient of determination is 0.47 between a dependent variable and an independent variable, This denotes that:

22 23 24

Select any one option.

[Clear A](#)

25 26 27

 The relationship between the two variables is not strong.

28 29 30

 The correlation coefficient between the two variables is also 0.47

31 32 33

 47% of the variance in the independent variable is explained by the dependent variable

34 35 36

 47% of the variance in the dependent variable is explained by the independent variable.

37 38 39

40 41 42

43 44 45

46 47 48

49 50 51

52 53 54

55 56 57

58 59 60

61 62 63

177 min left

28. C2 Clustering

[Previous](#)[Next](#)

MCQs:

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

31 32 33

34 35 36

37 38 39

40 41 42

43 44 45

46 47 48

49 50 51

52 53 54

55 56 57

58 59 60

61 62 63

A client has approached you for a problem statement that requires the use of clustering. You decided to model the problem statement with hierarchical clustering. Consider datasets having 'n' data points.

Which of the following statements is true for the above problem statement?

Answer Options

Select any one option:

[Clear A](#)

- 'n*n' distance matrix should be calculated for the mentioned problem statement
- Initially 'n' clusters are formed for the mentioned problem statement
- The output for the problem statement above is a dendrogram
- All the above

177 min left

27. C2 Clustering

[Previous](#)[Next](#)

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

31 32 33

34 35 36

37 38 39

40 41 42

43 44 45

46 47 48

49 50 51

52 53 54

55 56 57

58 59 60

61 62 63

Which of the following is not true for Hopkins statistic?

Answer Options

Select any one option.

[Clear A](#)

- Hopkins statistic decides if the data is suitable for clustering or not
- Hopkins statistic lie between -1 and 1
- If the Hopkins statistic comes out to be 0, then the data is uniformly distributed
- If the Hopkins statistic comes out to be 1, then the data highly suitable for clustering

177 min left

25. C2 Logistic Regression

[Previous](#)[Next](#)

MCQs:

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

31 32 33

34 35 36

37 38 39

40 41 42

43 44 45

46 47 48

49 50 51

52 53 54

55 56 57

58 59 60

61 62 63

For a completely random binary classification model, what will be the area under the curve of the ROC graph?

Answer Options

Select any one option.

[Clear A](#) 0 0.25 0.5 1

179 min left

7. C2 Logistic Regression

[Previous](#)[Next](#)

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

31 32 33

34 35 36



Recall the telecom churn example. If the log odds for churn are equal to 0 for a customer, then that means -

- Select any one option [Clear Answer](#)
- There is no chance of the customer churning
- The probability of the customer churning is equal to the probability of the customer not churning
- The probability of the customer churning is very small compared to the probability of the customer not churning
- The probability of the customer churning is very large compared to the probability of the customer not churning

Score

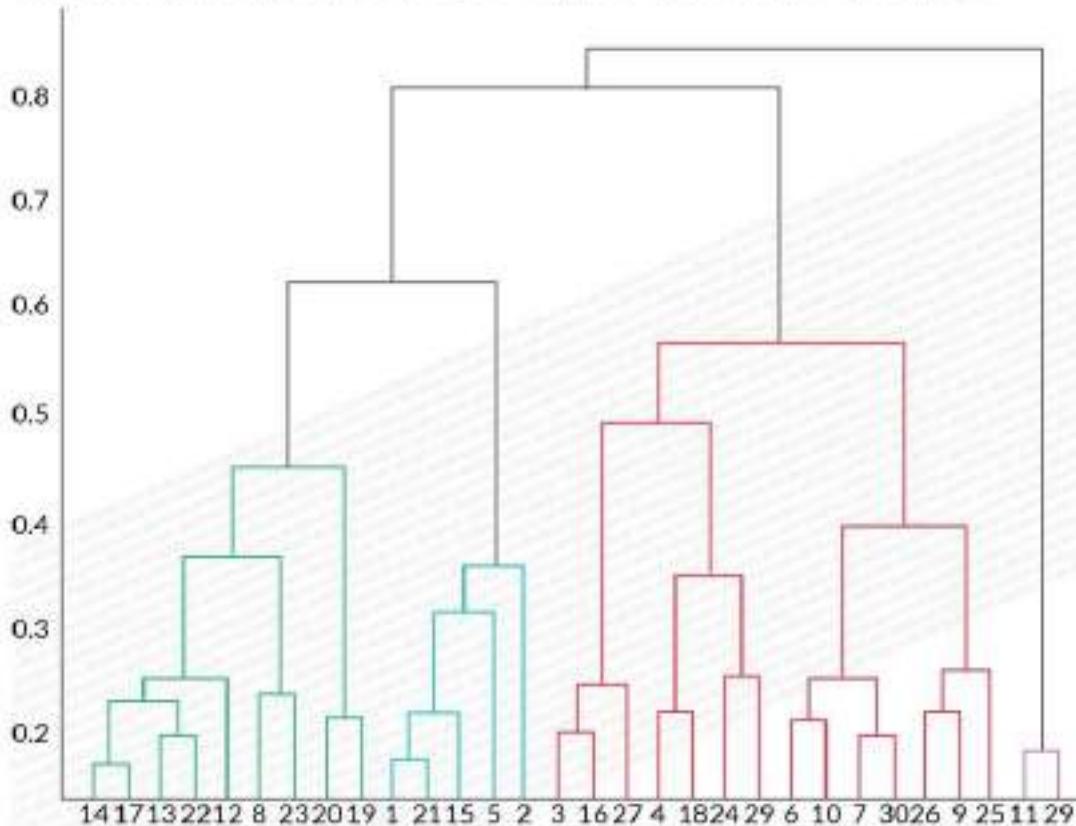
1

2

3

4

You observed the following dendrogram after performing k-means clustering on a dataset. Which of the following statements can be concluded from the dendrogram?



Answer Options

- Select one or more options:
- The final number of clusters is 6.
 - There are 23 data points used in the above clustering algorithm.
 - Single linkage is used to define the distance between two clusters in the above dendrogram.
 - The above dendrogram interpretation is not possible for K-Means clustering.

[Close Answer](#)

3. C2 linear Regression

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

31 32 33

34 35 36

37 38 39

40 41 42

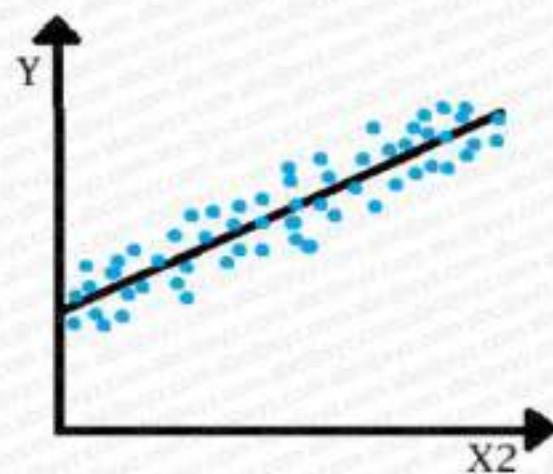
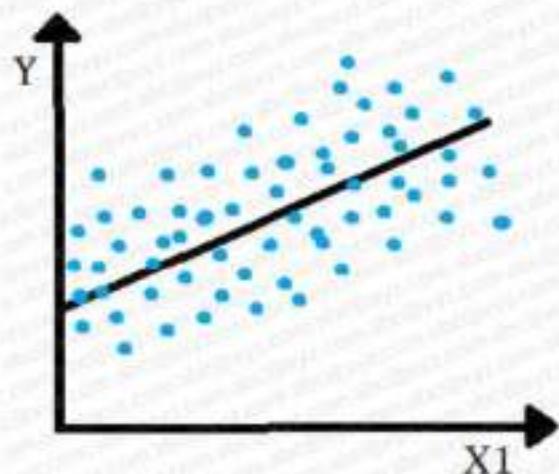
43 44 45

46 47 48

49 50 51

52 53 54

For the same dependent variable Y, two models were created using the independent variables X1 and X2. The following two graphs represent the fitted line on the scatterplot. (Both of the graphs are on the same scale)



Which of the following is true about the residuals in these two models?

Answer Options

Select any one option

Clear Answer

- The sum of residuals in model 2 is higher than model 1
- The sum of residuals in model 1 is higher than model 2
- Both have the same sum of residuals
- Nothing can be said about the sum of residuals from the graph



Submit Test

174 min left

69. C2 Multiple correct answer

[Previous](#)

Next

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

31 32 33

34 35 36

Which of the following command correctly builds a logistic regression model in Python? (More than one option may be correct.)

37 38 39

Answer Options

Select one or more options.

[Clear All](#)

from sklearn.linear_model import LogisticRegression
lr = LogisticRegression()
lr.fit(X_train, y_train)

import statsmodel.api as sm
lr = sm.GLM(y_train,(sm.add_constant(X_train)),
family = sm.families.Binomial())
lr.fit()

from sklearn.linear_model import LogisticRegression
lr = LogisticRegression()
lr.predict(X_train, y_train)

import statsmodel.api as sm
lr = sm.GLM(y_train,(sm.add_constant(X_train)),
family = sm.families.Binomial())
lr.predict()

Coding

1 2 3

4

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

31 32 33

34 35 36

37 38 39

40 41 42

43 44 45

46 47 48

49 50 51

52 53 54

55 56 57

58 59 60

61 62 63

64 65 66

67 68 69

70

Choose the correct option from the following.
The difference between '+' and '*' quantifier is _____.

Answer Options

Select any one option

Clear All

- '+' needs the preceding character to be present at least once whereas '*' does not need the same.
- '*' needs the character to be present at least once whereas '+' does not need the same.
- Both the quantifiers have the same functionality.
- None of the above.



Coding

1 2 3

4

174 min left

67. C2 linear Regression

[Previous](#)[Next](#)

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

31 32 33

34 35 36

37 38 39

40 41 42

43 44 45

46 47 48

49 50 51

52 53 54

55 56 57

58 59 60

61 62 63

64 65 66

67 68 69

70

Suppose you run a regression with one of the feature variables T, with all the remaining feature variables. The R-squared of this model was found out to be 0.8. What will be the VIF for variable T?

Answer Options

Select any one option

[Clear All](#) 1.56 2.77 3.33 5.00

67

Coding

1 2 3

4

174 min left

65. C2-Decision Trees

[Previous](#)[Next](#)

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

31 32 33

34 35 36

37 38 39

40 41 42

43 44 45

46 47 48

49 50 51

52 53 54

55 56 57

58 59 60

61 62 63

64 65 66

67 68 69

70

Suppose you train a decision tree with the following data. Which feature should we split on at the root?

X	Y	Z	V
T	T	F	1
F	F	F	0
T	T	T	0
F	T	T	1

Answer Options

Select one or more options

[Clear All](#) X Y Z Cannot be determined

Coding

1 2 3

4

174 min left

64. C2-Basics of NLP and Text Mining

[Previous](#)[Next](#)

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

31 32 33

34 35 36

37 38 39

Answer Options

40 41 42

Select any one option

[Clear All](#) 3 4 5 6

49 50 51

52 53 54

55 56 57

58 59 60

61 62 63

64 65 66

67 68 69

70

Coding

1 2 3

4

174 min left

63. C2 Multiple correct answer

[Previous](#)[Next](#)

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

31 32 33

34 35 36

37 38 39

40 41 42

43 44 45

46 47 48

49 50 51

52 53 54

55 56 57

58 59 60

61 62 63

64 65 66

67 68 69

70

Which of the following is NOT a methodology by which you can identify the optimal number of clusters for K-means clustering? (More than one option may be correct)

Answer Options

Select one or more options

[Clear All](#) Dendrogram inspection method Elbow Method Single Linkage Method Silhouette score

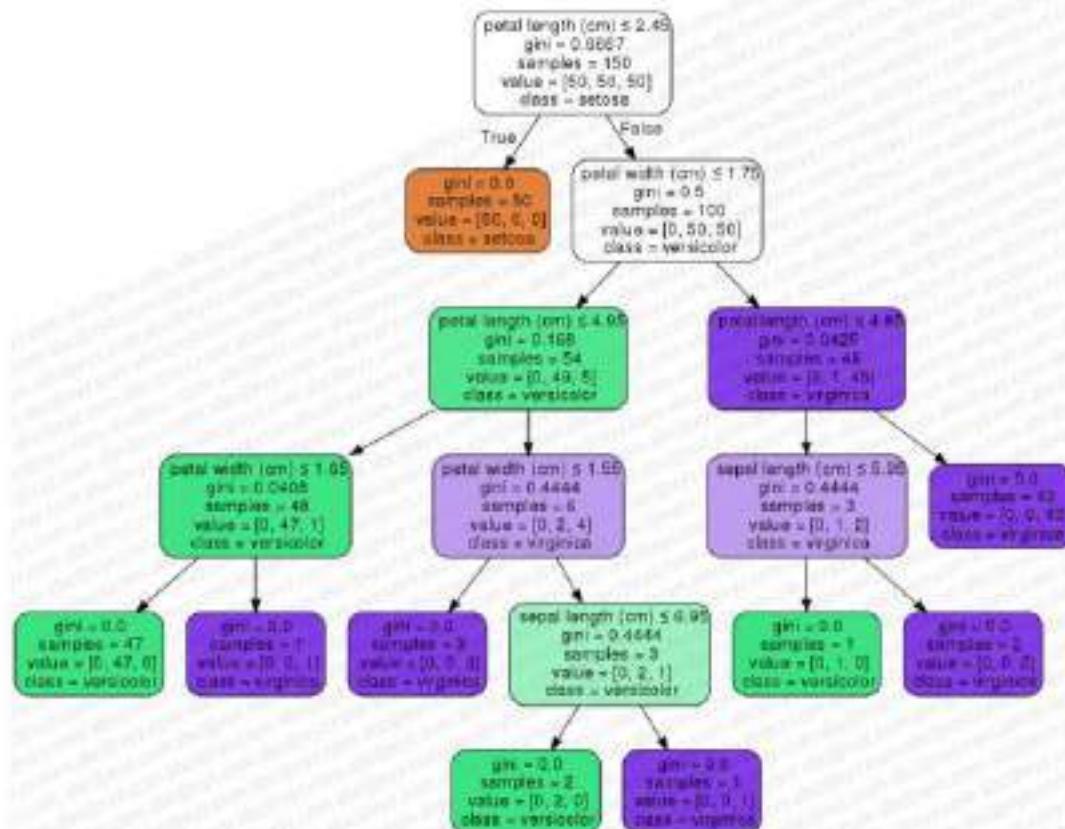
Coding

1 2 3

4

1	2	3
4	5	6
7	8	9
10	11	12
13	14	15
16	17	18
19	20	21
22	23	24
25	26	27
28	29	30
31	32	33
34	35	36
37	38	39
40	41	42
43	44	45
46	47	48
49	50	51
52	53	54
55	56	57
58	59	60
61	62	63
64	65	66
67	68	69
70		

Refer to the decision tree given below and choose the statement that is correct as per this tree.



Answer Options

Select any one option

- The tree given above will show very good performance on the train data.
- The tree given above is an underfitting tree.
- If the petal length is more than 2.45, then it is equally likely that the flower is either setosa or virginica.

175 min left

53. C2-Decision Trees

< Previous

Next

25 26 27

28 29 30

31 32 33

34 35 36

37 38 39

40 41 42

43 44 45

46 47 48

49 50 51

52 53 54

55 56 57

58 59 60

61 62 63

64 65 66

67 68 69

70

Coding

1 2 3

Which of the following metrics measures how often a randomly chosen element would be incorrectly identified?

Answer Options

Select any one option

Clear Ans

 Entropy Information Gain Gini Index None of these

175 min left

50. C2 Logistic Regression

[Previous](#)[Next](#)

25 26 27

28 29 30

31 32 33

34 35 36

37 38 39

40 41 42

43 44 45

46 47 48

49 50 51

52 53 54

55 56 57

58 59 60

61 62 63

64 65 66

67 68 69

70

Coding

1 2 3

You have built a Logistic Regression model that is trying to predict whether a loan is approved or not based on a person's FICO score. Here are the model parameters: Intercept (B_0) = -9.346 and coefficient of FICO score = 0.0146. Given these parameters, can you calculate the probability of a loan getting approved for someone with a FICO score of 655?

Answer Options

Select any one option

[Clear All](#) 0.35 0.45 0.55 0.65

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

Which of the following strings will match with the regular expression '^01+0\$'?

- 1. 0
- 2. 00
- 3. 0111111110

Answer Options

Select any one option

Clear Ans

Only option 1

Only option 3

Both 1&2

Both 2&3



40 41 42

43 44 45

46 47 48

49 50 51

52 53 54

176 min left

38. C2 linear Regression

[Previous](#)[Next](#)

MCQs:

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

What does standardised scaling do?

22 23 24

Answer Options

25 26 27

Select any one option.

[Clear A](#) Bring all data points in the range 0 to 1 Bring all data points in the range -1 to 1 Bring all the data points in a normal distribution with mean 0 and standard deviation 1 Bring all the data points in a normal distribution with mean 1 and standard deviation 0

37 38 39

40 41 42

43 44 45

46 47 48

49 50 51

52 53 54

55 56 57

58 59 60

61 62 63

176 min left

37. C2 Clustering

[Previous](#)[Next](#)

MCQs:

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

Which of the following is not true for Hopkins statistic?

22 23 24

Answer Options

25 26 27

Select any one option.

[Clear A](#) Hopkins statistic decides if the data is suitable for clustering or not

28 29 30

 Hopkins statistic lie between -1 and 1

31 32 33

34 35 36

 If the Hopkins statistic comes out to be 0, then the data is uniformly distributed

37 38 39

40 41 42

 If the Hopkins statistic comes out to be 1, then the data highly suitable for clustering

43 44 45

46 47 48

49 50 51

52 53 54

55 56 57

58 59 60

61 62 63

177 min left

35. C2 Logistic Regression

[Previous](#)[Next](#)

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

Consider the following confusion matrix.

Total=500	Actual Positive	Actual Negative
Predicted Positive	196	20
Predicted Negative	28	296

Which among the following is the lowest for the given confusion matrix?

Answer Options

Select any one option

[Clear Ans](#) Accuracy Precision Sensitivity Specificity

31 32 33

34 35 36

37 38 39

40 41 42

43 44 45

46 47 48

49 50 51

52 53 54

177 min left

33. C2 Logistic Regression

[Previous](#)[Next](#)

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

Consider the following univariate logistic model:

$$Y = \beta_0 + \beta_1 X_1$$

Which of the following statements is NOT true?

Answer Options

Select any one option

[Clear All](#) The maximum likelihood estimation determines the best combination of β_0 and β_1 . If β_1 is increased by 1 unit, Y increases by 1 unit. β_0 is the y-intercept If β_1 is increased by 1 unit, log-odds increases by 1 unit.

33

34 35 36

37 38 39

40 41 42

43 44 45

46 47 48

49 50 51

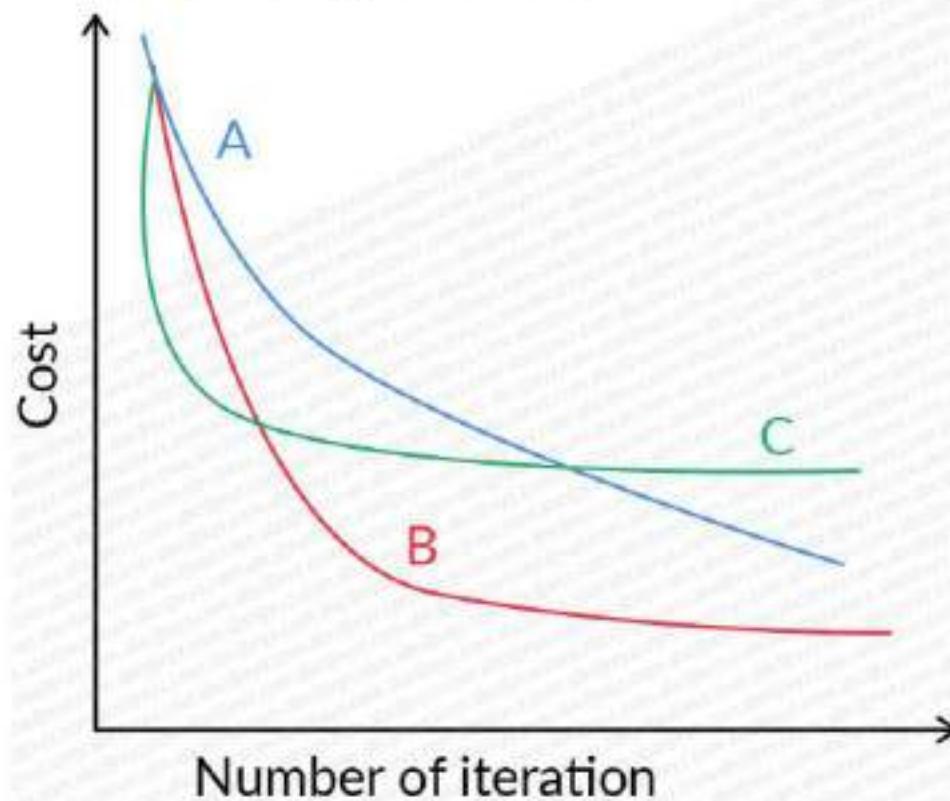
52 53 54

MCQs

- 1 2 3
- 4 5 6
- 7 8 9
- 10 11 12
- 13 14 15
- 16 17 18
- 19 20 21
- 22 23 24
- 25 26 27

- 28 29 30
- 31 32 33
- 34 35 36
- 37 38 39
- 40 41 42
- 43 44 45
- 46 47 48
- 49 50 51
- 52 53 54
- 55 56 57
- 58 59 60
- 61 62 63
- 64 65 66
- 67 68 69
- 70

Observe the following cost function graph with different learning rates.



Which of the following statements are true about the learning rate? (More than one option may be correct.)

Answer Options

Select one or more options

Clear

The learning rate of curve C is highest among all curves

The learning rate for curve B is lower than A

The learning rate for curve B is higher than A

28. C2 Logistic Regression

[Previous](#)[Next](#)

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

<<

25 26 27

28 29 30

31 32 33

34 35 36

37 38 39

30 31 32

Given an imbalanced dataset, the ratio of positive to negative class is 1:10000. You run a logistic regression model and find out that the model has a high value of precision and a low value of recall. Which of the following statements is true?

Answer Options

Select any one option

[Clear Ans](#)

- The class is handled well by the data
- The model is not able to detect the class, but when it does it is highly trustable
- The model is able to detect the class but it includes data points from the other class as well
- The class is handled poorly by the data

177 min left

27. C2 Clustering

[Previous](#)[Next](#)

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

<<

25 26 27

27

28 29 30

31 32 33

34 35 36

37 38 39

35 36 37

Initialising the following command in Python will result in the following: `model_clus = KMeans(n_clusters = 6, max_iter=50)`

Answer Options

Select any one option

[Clear Ans](#)

- Run maximum 6 iterations
- Run maximum 40 iterations
- Create 6 final clusters
- Create 50 final clusters

177 min left

26. C2 Logistic Regression

[Previous](#)[Next](#)

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

<<

25 26 27

28 29 30

31 32 33

34 35 36

37 38 39

30 31 32

Which of the following is correct for a logistic regression model?

Answer Options

Select any one option

[Clear Ans](#)

- The independent variables should not be multicollinear.
- The dependent variable should follow Normal Distribution.
- The log odds in a logistic regression model lies between 0 and 1.
- F1-score is always the best metric for evaluating a logistic regression model.

177 min left

25. C2 linear Regression

[Previous](#)[Next](#)

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

<<

25 26 27

28 29 30

31 32 33

34 35 36

37 38 39

Which of the following is true regarding the error terms in linear regression?

Answer Options

Select any one option

[Clear Ans](#)

- The sum of residuals should be zero
- The sum of residuals should be lesser than zero
- The sum of residuals should be greater than zero
- There is no such restriction on what the sum of residuals should be

24. C2 Logistic Regression

[Previous](#)[Next](#)

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

Which of the following is true for weight of evidence (WoE) analysis?

[Clear Ans](#)

Answer Options

Select any one option

- It helps in finding the different predictive patterns for the different segments that might be present in the data.
- WoE helps in treating missing values for both continuous and categorical variables.
- WoE values should follow an increasing or decreasing trend across bins.
- All of the above

35 36

37 38 39

35 36

37 38 39

35 36

37 38 39

23. C2 Multiple correct answer

[◀ Previous](#)[Next ▶](#)

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

<<

25 26 27

28 29 30

31 32 33

34 35 36

37 38 39

In a simple linear regression model when you fit a straight line through the data you'll get the two parameters of the straight line, i.e. the intercept β_0 and the slope β_1 . Which of the following is true for β_0 and β_1 ? (More than one option may be correct.)

Answer Options

Select one or more options

[Clear Ans](#) The null hypothesis for a simple linear regression model is $H_0: \beta_1 = 0$ If the p-value turns out to be greater than 0.05 for β_1 , it means β_1 is significant If β_1 turns out to be insignificant, that means there is no relationship between the dependent and independent variable If the p-value turns out to be less than 0.05 for β_0 , it means that β_0 is non-zero

20. C2 linear Regression

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

<<

25 26 27

28 29 30

31 32 33

34 35 36

37 38 39

40 41 42

In linear regression, the metric F-statistic is used to determine

Clear Ans

 the significance of the individual beta coefficients the variance explanation strength of the model the significance of the overall model fit Both A & C

18. C2 linear Regression

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

<<

25 26 27

28 29 30

31 32 33

34 35 36

37 38 39

40 41 42

Consider the following two assumptions for a simple regression model. (Assume X and y to be independent and dependent variables respectively).

Statement 1: There is a linear relationship between X and y.

Statement 2: X and y are normally distributed.

Answer Options

Select any one option

Clear Ans

 Statement 1 is correct and Statement 2 is incorrect Statement 1 is incorrect and Statement 2 is correct Both the statements are correct Both the statements are incorrect

178 min left

17. C2 Clustering

[◀ Previous](#)[Next ▶](#)

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

<<

25 26 27

28 29 30

31 32 33

34 35 36

37 38 39

39 40 41

Which of the following statements is NOT true?

Answer Options

Select any one option

[Clear Ans](#)

- Each time the clusters are made during the K-means algorithm, the centroid is updated.
- The cluster centres that are computed in the K-means algorithm are given by centroid value of the cluster points.
- Standardization of the data is not important before applying Euclidean distance as a measure of similarity/dissimilarity
- The centroid of a column with data points 25, 32, 34 and 23 is 28.5

178 min left

16. C2 Logistic Regression

[◀ Previous](#)[Next ▶](#)

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

<<

25 26 27

28 29 30

31 32 33

34 35 36

37 38 39

40 41 42

If you use a random number generator to predict the output 0 or 1 for a binary classification problem, what will be the area under the curve of the ROC curve?

Answer Options

Select any one option

[Clear Ans](#) 0 0.5 1 100

15. C2 Business Problem Solving

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

<<

 Neural Network

25 26 27

 Logistic regression

28 29 30

 Decision tree

31 32 33

 All of the above

34 35 36

37 38 39

The coronavirus disease (COVID-19), was declared a pandemic by the World Health Organization (WHO) in February 2020. Currently, there are no vaccines or treatments that have been officially approved by WHO after clinical trials. India has not seen the peak of infection yet and the number of infections is touching a new height daily. The business unit of an Indian health and hygiene company approaches you to know "Why the sales of masks is decreasing despite the number of corona infections increasing daily".

Answer the following questions:

Suppose you mapped the above problem statement with a classification problem, either a customer will buy a mask or not. You will build _____ model as your initial solution.

Answer Options

Select any one option

Clear Ans

 Neural Network Logistic regression Decision tree All of the above

12. C2 Business Problem Solving

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

The coronavirus disease (COVID-19), was declared a pandemic by the World Health Organization (WHO) in February 2020. Currently, there are no vaccines or treatments that have been officially approved by WHO after clinical trials. India has not seen the peak of infection yet and the number of infections is touching a new height daily. The business unit of an Indian health and hygiene company approaches you to know "**Why the sales of masks is decreasing despite the number of coronaviruses increasing daily?**".

Answer the below question:

Consider the following two statements:

Statement 1: Understanding the change in customer behaviour is an important factor to be considered for business understanding for the problem statement above.

Statement 2: One of the possible hypotheses for the above problem statement: There is a rise in the number of companies manufacturing normal/surgical masks due to which the sales of the client's company is decreasing.

Answer Options

<<

Select any one option

Clear Ans

 Statement 1 is correct and Statement 2 is wrong Statement 2 is correct and Statement 1 is wrong Both the statements are correct

178 min left

11. C2 linear Regression

[Previous](#)[Next](#)

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

<<

25 26 27

28 29 30

31 32 33

34 35 36

37 38 39

How is regression different from classification?

Answer Options

Select any one option

[Clear Ans](#)

- One is supervised while the other is unsupervised
- One is iterative while the other is closed form
- In regression, the response variable is numeric while it is categorical in classification
- None of the above

10. C2 Business Problem Solving

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24



Answer Options

25 26 27

Select any one option

Clear Ans

28 29 30

 3>1>5>2>4>7>6>8

31 32 33

 3>2>1>5>4>7>6>8

34 35 36

 4>3>1>2>5>7>6>8

37 38 39

 3>2>1>5>4>7>8>6

40 41 42



9. C2 Multiple correct answer

[◀ Previous](#)[Next ▶](#)

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

<<

25 26 27

28 29 30

31 32 33

34 35 36

37 38 39

ROC curve shows the tradeoff between the True Positive Rate (TPR) and the False Positive Rate (FPR). TPR and FPR are sensitivity and (1 - specificity) respectively. The following function is written in Python using metrics package from the scikit-learn library for a ROC curve function.

```
def draw_roc(actual, probs):
    fpr tpr,thresholds=metrics.roc_curve(actual,probs,drop_intermediate=False)
    auc_score = metrics.roc_auc_score(actual, probs)
    return None
```

Which of the following statements are true? (More than one option may be correct.)

Answer Options

Select one or more options:

[Clear Ans](#) The area under the ROC curve can be more than 1. The arguments passed in the above function are actual values of the target variable and the predicted values (i.e., 0 or 1) Larger the area under the curve, the better will be the model

174 min left

37. C3 Multiple Correct Answer

[Previous](#)[Next](#)

Course 2

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

How is diversity achieved in the case of Random Forest? (More than one option may be correct.)

Answer Options

Select one or more options

[Clear Ans](#)

- Bootstrapped sampling is performed over the data to create multiple samples.
- We perform tree pruning for each tree built on the bootstrapped sample.
- A random subset of features are considered at each node of the tree built on the bootstrapped sample.
- A random subset of features are considered for building each tree.

Course 3

1 2 3

174 min left

36. C3- Advanced ML

[Previous](#)[Next](#)

Course 2

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27



28 29 30

31 32 33

34 35 36

37

Course 3

1 2 3

Suppose an imbalanced data set has a class ratio of 2:3, and you want to run a cross-validation scheme to evaluate a model's performance. If you apply a stratified k-fold to generate the train-test folds, what will be the distribution of the classes in the test split?

Answer Options

Select any one option

[Clear Ans](#) 1:5 2:3 1:7 None of these

179 min left

6. C2 linear Regression

[Previous](#)[Next](#)

Course 2

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

What does standardised scaling do?

[Clear Ans](#)

Answer Options

Select any one option

- Bring all data points in the range 0 to 1
- Bring all data points in the range -1 to 1
- Bring all the data points in a normal distribution with mean 0 and standard deviation 1
- Bring all the data points in a normal distribution with mean 1 and standard deviation 0

<<

37

Course 3

1 2 3

35. C3 Time Series Forecasting

Course 2

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

Consider the following two problem statements:

Problem statement A: You have to predict the price of a stock for the next day based on the stock prices of the previous four days.

Problem statement B: You have to predict the price of a stock for the next month based on monthly data of the last 5 years. The stock price has an increasing trend.

Which of these forecasting methods are correctly mapped with the problem statements?

Answer Options

Select any one option

Clear Ans

 A: Simple moving average, B: Holt's Method A: Naive method, B: Holt-Winters' method A: Simple moving average, B: Naive method A: SARIMA, B: SARIMAX

Course 3

1 2 3

174 min left

34. C3 Time Series Forecasting

[Previous](#)[Next](#)

Course 2

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

Answer Options

Select any one option

[Clear Ans](#) 6 5 3 None of these

37

Course 3

1 2 3

33. C3 Time Series Forecasting

Course 2

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8
- 9
- 10
- 11
- 12
- 13
- 14
- 15
- 16
- 17
- 18
- 19
- 20
- 21

- 22
- 23
- 24

- 25
- 26
- 27



Clear Ans

- 28
- 29
- 30

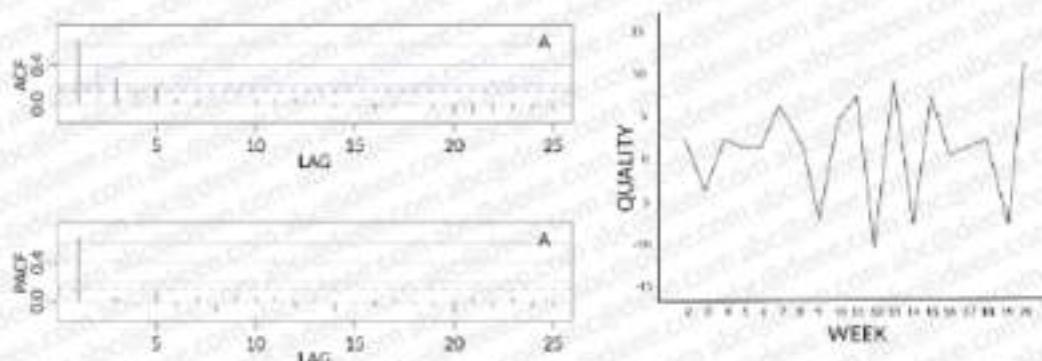
- 31
- 32
- 33

- 34
- 35
- 36

- 37

Course 3

- 1
- 2
- 3



Answer Options

Select any one option

 (1,0,1) (1,1,5) (1,2,5) (1,1,1)

Course 2

- 1
- 2
- 3

- 4
- 5
- 6

- 7
- 8
- 9

- 10
- 11
- 12

- 13
- 14
- 15

- 16
- 17
- 18

- 19
- 20
- 21

- 22
- 23
- 24

- 25
- 26
- 27

- 28
- 29
- 30

- 31
- 32
- 33

- 34
- 35
- 36

- 37



Course 3

- 1
- 2
- 3

- 4
- 5
- 6

- 7
- 8
- 9

- 10
- 11
- 12

- 13
- 14
- 15

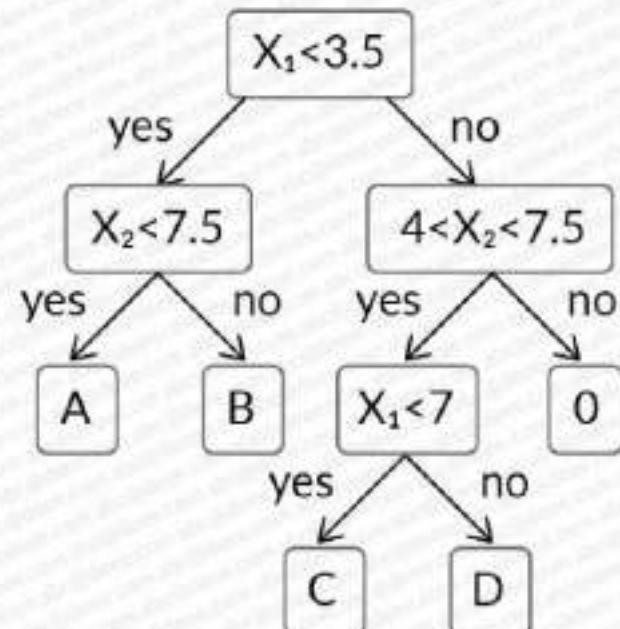
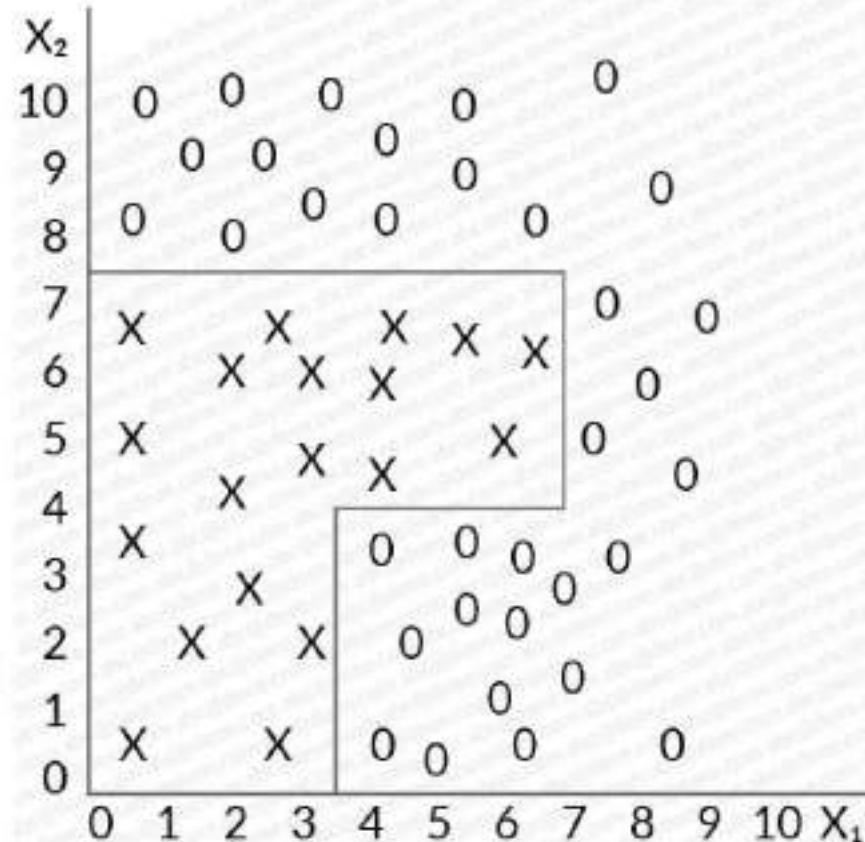
- 16
- 17
- 18

- 19
- 20
- 21

- 22
- 23
- 24

Refer to the image with the decision tree and the boundary diagram. Find what will be the outcome for the leaf nodes A, B, C, D.

Note: An outcome can be an "o" or an "x"



Answer Options

Select any one option

Close A

 A: x, B: x, C: o, D: o A: o, B: o, C: x, D: x

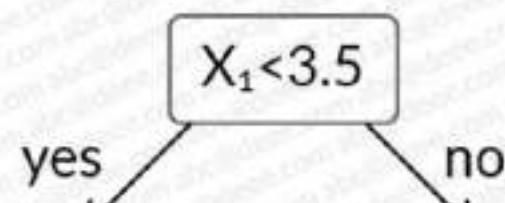
Course 2

- | | | |
|----|----|----|
| 1 | 2 | 3 |
| 4 | 5 | 6 |
| 7 | 8 | 9 |
| 10 | 11 | 12 |
| 13 | 14 | 15 |
| 16 | 17 | 18 |
| 19 | 20 | 21 |
| 22 | 23 | 24 |
| 25 | 26 | 27 |
| 28 | 29 | 30 |
| 31 | 32 | 33 |
| 34 | 35 | 36 |
| 37 | | |

Refer to the image with the decision tree and the boundary diagram. Find what will be the outcome for the leaf node: A, B, C, D.

Note: An outcome can be an "o" or an "x"

X_2	10	9					
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0



Course 3

- | | | |
|---|---|---|
| 1 | 2 | 3 |
|---|---|---|

Answer Options

32. C3 Advanced ML

[Previous](#)[Next](#)

Course 2

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

<<

28 29 30

31 32 33

34 35 36

37

Course 3

1 2 3

Refer to the image with the decision tree and the boundary diagram. Find what will be the outcome for the leaf node: A, B, C, D.

Note: An outcome can be an "o" or an "x"

Answer Options

Select any one option

[Clear Ans](#) A: x, B: x, C: o, D: o A: o, B: o, C: x, D: x A: x, B: o, C: x, D: o A: o, B: x, C: o, D: x

174 min left

31. C3 Time Series Forecasting

[Previous](#)[Next](#)

Course 2

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

<<

28 29 30

31 32 33

34 35 36

37

Course 3

1 2 3

The actual observation and simple exponential forecasted value for period t-1 were 120 and 115 respectively. What is the SES forecast for period t when $\alpha = 0.3$?

Answer Options

Select any one option

[Clear Ans](#) 115.5 116 116.5 117

30. C3- Advanced ML

[Previous](#)[Next](#)

Course 2

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27



28 29 30

31 32 33

34 35 36

37

Course 3

1 2 3

Feature engineering is an important step in any model building exercise. It is the process of creating new features from a given data set using the domain knowledge to leverage the predictive power of a machine learning model. Which of the following statements are correct?

Statement 1: Feature engineering techniques are applied before train-test split.

Statement 2: There is no difference between standardization and normalization.

Statement 3: Mean encoding is a feature engineering technique for handling categorical features.

Answer Options

Select any one option

[Clear Ans](#) Only 1 and 2 Only 2 and 3 Only 1 Only 3

Course 2

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

Answer Options

Select any one option

Clear Ans

 0.137 0.267 0.527 0.397

Course 3

1 2 3

Random Forest			
Tree Number	Gini Impurity before split	Gini Impurity after split	
1	0.39	0.31	
2	0.46	0.28	
3	0.40	0.21	
4	0.42	0.32	

175 min left

28. GroupBy

< Previous

Next >

SOLVE

1. -- Enter your query here

Course 2

1	2	3
4	5	6
7	8	9
10	11	12
13	14	15
16	17	18
19	20	21
22	23	24
25	26	27
28	29	30
31	32	33
34	35	36

INPUT TABLE

You're given a **orders** table and the columns in the orders table are shown below:

orders		
Order_Id	INT	
Type	VARCHAR(10)	
Real_Shipping_Days	INT	
Scheduled_Shipping_Days	INT	
Customer_Id	INT	
Order_City	VARCHAR(20)	
Order_Date	DATE	
Order_Region	VARCHAR(15)	
Order_State	VARCHAR(20)	
Order_Status	VARCHAR(20)	
Shipping_Mode	VARCHAR(20)	

Course 3

1	2	3
4	5	6
7	8	9
10	11	12
13	14	15
16	17	18
19	20	21
22	23	24

QUERY

- Calculate count of all the orders.
- Where Order_State is Maharashtra
 - Note** - Use the alias of **oc** for count of orders.
- Group the results by **Type**.
- Order them by **oc** in ascending order.

OUTPUT COLUMNS

oc, Type

Here's an image showing how a sample output would look like:

OC	TYPE
45	CASH
89	PAYMENT
108	TRANSFER
151	DEBIT

Note that the coding console automatically converts the casing of the columns to upper case

Console

Custom Test Case

180 min left

5. C2 Logistic Regression

[Previous](#)[Next](#)

Course 2

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

Which of the following is true for weight of evidence (WoE) analysis?

Answer Options

Select any one option

[Clear Ans](#) It helps in finding the different predictive patterns for the different segments that might be present in the data. WoE helps in treating missing values for both continuous and categorical variables. WoE values should follow an increasing or decreasing trend across bins. All of the above

Course 3

1 2 3

175 min left

27. C3 Time Series Forecasting

[◀ Previous](#)[Next ▶](#)

Course 2

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

Answer Options

Select any one option

Clear Ans

 3 2 1 0

Course 3

1 2 3

4 5 6

7 8 9

15 16 17

175 min left

26. C3 Advanced ML

[Previous](#)[Next](#)

Course 2

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30



31 32 33

34 35 36

37

Course 3

1 2 3

4 5 6

7 8 9

Refer to the image and choose the best option that represents box A and B.

Answer Options

Select any one option

[Clear All](#) A: Underfitting, B: Good Model A: Good Model, B: Overfitting A: Underfitting, B: Overfitting A: Overfitting, B: Underfitting

175 min left

25. C3 Advanced ML

[Previous](#)[Next](#)

Course 2

Which of the following describes an advantage of the random forest (RF) algorithm over the decision tree (DT) algorithm?

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

Answer Options

19 20 21

Select any one option

[Clear Ans](#)

22 23 24

 An RF model has better interpretability as compared to a DT

25 26 27

 An optimally fit RF model has less variance than an optimally fit DT

28 29 30



31 32 33

 An optimally fit RF model has more variance than an optimally fit DT

34 35 36

37

 Building an RF model is computationally less expensive than building a DT

1 2 3

4 5 6

7 8 9

10 11 12

Course 2

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30



31 32 33

34 35 36

37

Course 3

1 2 3

4 5 6

7 8 9

Refer to the table below and find the OOB Error for a Random Forest model that consists of 4 trees namely A, B, C, D and 10 observations. Use the table below and fill the column "Predicted Label" and then make use of it to calculate the OOB error.

Class Labels:

"0": Positive Class

"-": Negative Class

Observations	Trees Predictions				Actual Label	Predicted Label
	A	B	C	D		
3	-	-	-	0	-	-
10	0	0	0	-	0	0
5	-	-	0	-	0	0
2	-	-	-	0	-	-
8	0	0	0	-	0	0
1	0	-	0	0	-	-
6	-	-	-	-	-	-
9	0	0	0	0	0	0
4	-	-	-	0	0	0
7	0	0	0	-	0	0

Answer Options

Select any one option

Clear Ans

 0.1 0.2

Course 2

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8
- 9
- 10
- 11
- 12
- 13
- 14
- 15
- 16
- 17
- 18
- 19
- 20
- 21

What would happen if you choose a very high value of the hyperparameter, lambda?

$$\min_{a,b} \left[\sum_{i=1}^n (y_i - ax_i - b)^2 + \lambda(a^2 + b^2) \right]$$

- 22
- 23
- 24
- 25
- 26
- 27

Answer Options

Select any one option

Clear Ans

- The model would become simpler, yet it would show robust performance on test data.
- The model would become too simple. It would become an underfitted model.
- The model would become too complex. It would become an overfitted model.
- The model would show good performance on the train data, but it will show poor performance on the test data.

Course 3

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8
- 9
- 10
- 11
- 12

175 min left

22. C3 Advanced ML

[Previous](#)[Next](#)

Course 2

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

Decision Tree is a high variance model. Which of the following correctly explains this statement?

19 20 21

Select any one option

[Clear Ans](#)

22 23 24

 The number of attributes on both the sides of the decision tree is not the same.

25 26 27

 The decision tree building process is a top-down approach.

28 29 30

 The entire structure of the tree changes with small variations in the input data.

31 32 33

 All the above

34 35 36

37

Course 3

1 2 3

4 5 6

7 8 9

10 11 12

175 min left

21. C3 Time Series Forecasting

[◀ Previous](#)[Next ▶](#)

Course 2

Which is the compulsory component of any time series?

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

Answer Options

Select any one option

[Clear All](#) Level Trend Seasonality All of the above

37

Course 3

1 2 3

4 5 6

7 8 9

10 11 12

175 min left

20. C3 Advanced ML

[◀ Previous](#)[Next ▶](#)

Course 2

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

Which of the following methods is NOT true for the truncation of a decision tree?

Answer Options

Select any one option

[Clear Ans](#) Limit the depth of a tree Set a minimum threshold on the number of samples that appear in a leaf. Merging of two non-leaf nodes. Truncation is also known as pre-pruning.

Course 3

1 2 3

4 5 6

7 8 9

10 11 12

175 min left

19. C3- Advanced ML

[Previous](#)[Next](#)

Course 2

- 1 2 3
- 4 5 6
- 7 8 9
- 10 11 12
- 13 14 15

- 16 17 18

- 19 20 21

- 22 23 24

- 25 26 27

- 28 29 30

- 31 32 33

- 34 35 36

- 37

Course 3

- 1 2 3
- 4 5 6
- 7 8 9

What will be the accuracy percentage of the given confusion matrix of the three-class classification?

True/ Predicted	Class A	Class B	Class C
Class A	13	0	5
Class B	0	4	8
Class C	1	1	9

Answer Options

Select any one option

[Clear Ans](#) 63% 36% 71% 45%

175 min left

18. C3 Multiple Correct Answer

< Previous

Next

Course 2

- 1
- 2
- 3

- 4
- 5
- 6

- 7
- 8
- 9

- 10
- 11
- 12

- 13
- 14
- 15

- 16
- 17
- 18

- 19
- 20
- 21

- 22
- 23
- 24

- 25
- 26
- 27

- 28
- 29
- 30

- 31
- 32
- 33

- 34
- 35
- 36

- 37

Consider the following confusion matrix that summarises the prediction made by a model.

			Predicted	
			Achieved = YES	Achieved = No
Actual	Achieved = YES	100(TP)	40(FN)	
	Achieved = No	50(FP)	60(TN)	

Which of the following option(s) is/are correct?

Answer Options

Select one or more options

Clear All

 Accuracy is ~0.72

 Precision is ~0.67

 Recall is ~0.7

 F1 score is ~0.69

Course 3

- 1
- 2
- 3

- 4
- 5
- 6

- 7
- 8
- 9

- 10
- 11
- 12

180 min left

4. C2 Logistic Regression

[Previous](#)[Next](#)

Course 2

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

<<

28 29 30

31 32 33

34 35 36

37

If you use a random number generator to predict the output 0 or 1 for a binary classification problem, what will be the area under the curve of the ROC curve?

Answer Options

Select any one option

[Clear Ans](#) 0 0.5 1 100

175 min left

18. C3 Multiple Correct Answer

[Previous](#)[Next](#)

Course 2

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

Consider the following confusion matrix that summarises the prediction made by a model.

Which of the following option(s) is/are correct?

Answer Options

Select one or more options

[Clear All](#) Accuracy is ~0.72 Precision is ~0.67 Recall is ~0.7 F1 score is ~0.69

Course 3

1 2 3

4 5 6

7 8 9

175 min left

17. C3 Advanced ML

[Previous](#)[Next](#)

Course 2

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

Which of the following statements is NOT true for the decision tree regression?

19 20 21

Answer Options

Select any one option

[Clear Ans](#)

22 23 24

 Leaves in decision tree regression contain average values as the prediction.

25 26 27

 Impurity measure for a given node is measured by the weighted mean square error.

28 29 30

 In decision tree regression, a lower value of mean square error means that the data values are dispersed widely around mean.

31 32 33

 Weighted mean square error is nothing but the variance of the observations.

34 35 36

37

Course 3

1 2 3

4 5 6

7 8 9

10 11 12

Course 2

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30



31 32 33

34 35 36

37

Course 3

1 2 3

4 5 6

7 8 9

10 11 12

Suppose you built a random forest classifier model. You observed that the accuracy of the model turns out to be 99% on the train data but performs badly on test data.

The senior data scientist of your company suggested you to create a model on the subsets of the training data and test the same model on different subsets of the training data.

Which of the following statements is NOT true considering the above scenario?

Answer Options

Select any one option

Clear All

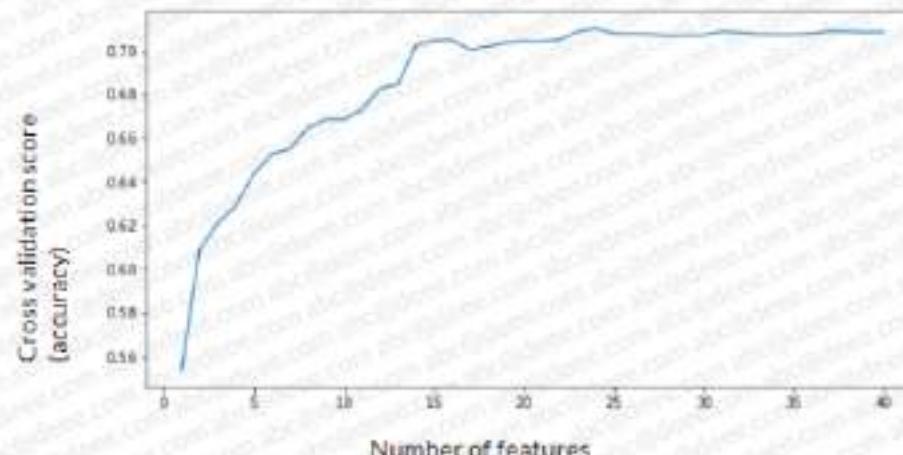
- The senior data scientist is talking about the Cross-validation scheme here.
- Cross-validation schemes can only give you reliable insights if your data is very large.
- OOB score is similar to cross-validation score in random forest.

- Cross-validation score is likely to be less than the accuracy of the model you built.

Course 2

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8
- 9
- 10
- 11
- 12
- 13
- 14
- 15
- 16
- 17
- 18
- 19
- 20
- 21

Observe the following graph for a logistic regression model. The graph determines the Cross-validation score based on the different number of features.



Which of the following statements is NOT true based on the above graph?

- 22
- 23
- 24
- 25
- 26
- 27
- 28
- 29
- 30
- 31
- 32
- 33
- 34
- 35
- 36
- 37



Answer Options

Select any one option

Clear Ans

- There are 40 features in the model.
- The performance of the model is lowest with one feature.
- The performance of the model is optimum with the number of features between 15 and 25.

Course 3

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8
- 9
- 10
- 11
- 12

175 min left

14. C3 Multiple Correct Answer

< Previous

Next

Course 2

- 1
- 2
- 3

- 4
- 5
- 6

- 7
- 8
- 9

- 10
- 11
- 12

- 13
- 14
- 15

- 16
- 17
- 18

- 19
- 20
- 21

- 22
- 23
- 24

- 25
- 26
- 27

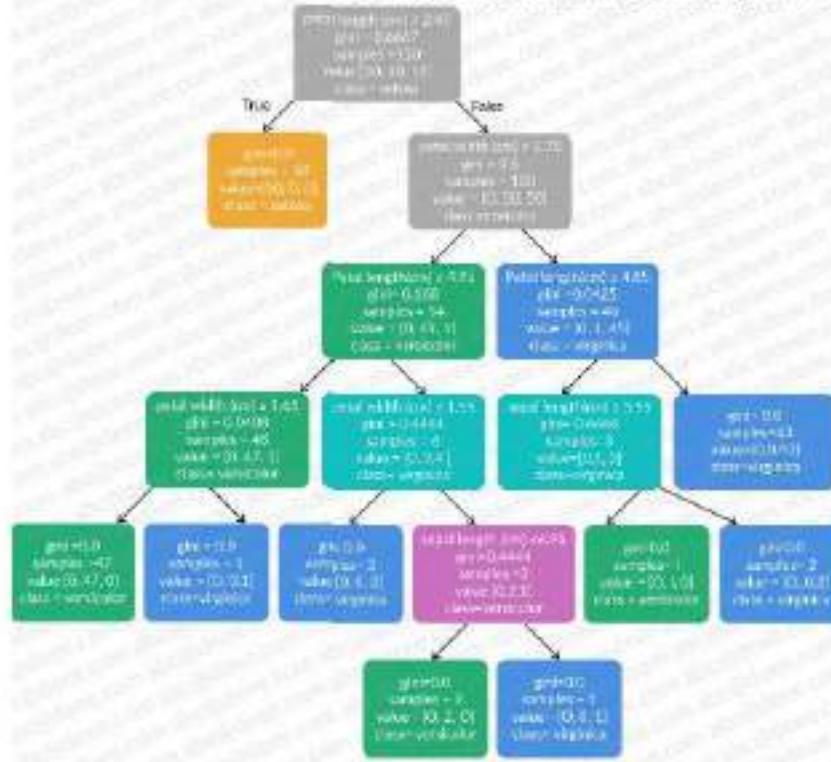
- 28
- 29
- 30

- 31
- 32
- 33

- 34
- 35
- 36

- 37

Refer to the decision tree and choose all the correct statements according to the decision tree provided. (More than one option may be correct.)



Course 3

Answer Options

Select one or more options.

Clear A

- The given tree is an overfitting tree.

- The given tree will be having a stable performance, if we change one row from the training data.

- If the petal length is more than 2.45 that it's equally likely that the flower is either a versicolor or virginica.

176 min left

13. C3 Advanced ML

[Previous](#)[Next](#)

34 35 36

37

Course 3

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

31 32 33

34 35 36

37

The following command is initialised in Python using the scikit-learn library for a decision tree model.

```
model=DecisionTreeClassifier(max_depth=4,min_samples_split=20,random_state=42)
```

Which of the following statements are true?

Answer Options

Select any one option

[Clear Ans](#)

- The homogeneity metric used here is Gini.
- The random state is passed to make the output decision tree consistent.
- The minimum number of samples required to split an internal node is 20.
- All of the above

176 min left

13. C3 Advanced ML

[Previous](#)[Next](#)

Course 2

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

The following command is initialised in Python using the scikit-learn library for a decision tree model.

```
model=DecisionTreeClassifier(max_depth=4,min_samples_split=20,random_state=42)
```

Which of the following statements are true?

Answer Options

Select any one option

[Clear Ans](#)

- The homogeneity metric used here is Gini.
- The random state is passed to make the output decision tree consistent.
- The minimum number of samples required to split an internal node is 20.
- All of the above

Course 3

1 2 3

176 min left

12. C3 Time Series Forecasting

[Previous](#)[Next](#)

Course 2

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

<<

28 29 30

31 32 33

34 35 36

37

Course 3

1 2 3

You are an analyst at a retail company in India. Owing to the COVID-19 pandemic, there is a huge demand for sanitisers, and the rate of infection is following an upward trajectory (a higher number of infections daily). You have access to past 30 days data for the demand of sanitisers. Which of the following models would work best for forecasting the demand for sanitisers for the next 5 days with highest accuracy?

Answer Options

Select any one option

[Clear Ans](#) Seasonal Auto-Regressive Integrated Moving Average Model Simple Average Forecast Method Simple Exponential Smoothing Method Holt-Winters' Method

Course 2

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

<<

28 29 30

31 32 33

34 35 36

37

Course 3

1 2 3

We wanted to tune hyperparameters for a Random Forest model using Grid Search technique with the CV of 5 folds. Refer to the list of hyperparameters that are required to be tuned.

```
{'n_estimators': [10, 25], 'max_features': [5,10],  
 'max_depth': [10, 50, None], 'bootstrap': [True, False]  
}
```

Assume that each set of hyperparameters takes 1 minutes to run, find the total time required to complete the tuning process.

Answer Options

Select any one option

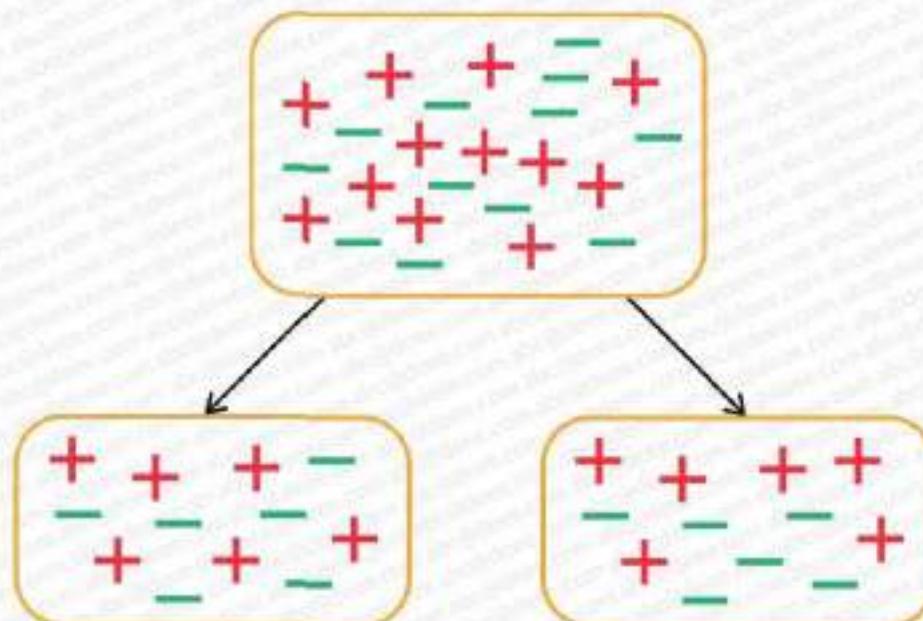
Clear Ans

 60 minutes 45 minutes 90 minutes 120 minutes

Course 2

- 1 2 3
- 4 5 6
- 7 8 9
- 10 11 12
- 13 14 15
- 16 17 18
- 19 20 21
- 22 23 24
- 25 26 27
- 28 29 30
- 31 32 33
- 34 35 36

Refer to the following decision tree.



Consider the following statements.

37

Statement 1: The given tree is not a decision tree as both the leaf nodes are heterogeneous.**Statement 2:** The split is done incorrectly. The leaf nodes are as impure as the root node.

Course 3

- 1 2 3
- 4 5 6
- 7 8 9
- 10 11 12
- 13 14 15
- 16 17 18

Answer Options

Select any one option

Clear A

 Statement 1 is correct and Statement 2 is incorrect Statement 1 is incorrect and Statement 2 is correct

3. C2 linear Regression

Course 2

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

<<

28 29 30

31 32 33

34 35 36

37

Course 3

1 2 3

Consider the following two assumptions for a simple regression model. (Assume X and y to be independent and dependent variables respectively).

Statement 1: There is a linear relationship between X and y.

Statement 2: X and y are normally distributed.

Answer Options

Select any one option

[Clear Ans](#) Statement 1 is correct and Statement 2 is incorrect Statement 1 is incorrect and Statement 2 is correct Both the statements are correct Both the statements are incorrect

176 min left

9. C3 Multiple Correct Answer

< Previous

Next

Course 2

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8
- 9
- 10
- 11
- 12
- 13
- 14
- 15
- 16
- 17
- 18
- 19
- 20
- 21

- 22
- 23
- 24
- 25
- 26
- 27
- 28
- 29
- 30
- 31
- 32
- 33
- 34
- 35
- 36
- 37

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8
- 9

Find the value of A and B using MA process given the value of SIGMA is 0.5 and MU is 15. (More than one option may be correct.)

Year	Forecast	Error	Actual
1990			15
1991			16
1992	A		13
1994		B	16
1995			14
1996			16

Answer Options

Select one or more options

Clear All

 A: 15.5 B: -1.25 B: 1.75 A: 16.5

176 min left

9. C3 Multiple Correct Answer

[Previous](#)[Next](#)

Course 2

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

Answer Options

Select one or more options

[Clear All](#) A: 15.5 B: -1.25 B: 1.75 A: 16.5

Course 3

1 2 3

4 5 6

7 8 9

176 min left

8. C3 Time Series Forecasting

[Previous](#)[Next](#)

Course 2

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30



31 32 33

34 35 36

37

Course 3

1 2 3

4 5 6

7 8 9

What does a MAPE value of 10% mean?

Answer Options

Select any one option

[Clear All](#) The absolute difference between the actual value and the forecasted value is 10%. The average absolute difference between the actual value and forecasted value is 10%. The forecasted value is 10% behind the actual value. The forecasted value is 10 times the actual value.

176 min left

7. C3 Multiple Correct Answer

[Previous](#)[Next](#)

Course 2

Which of the scenario(s) is/are a time series problem? (More than one options may be correct)

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8
- 9
- 10
- 11
- 12
- 13
- 14
- 15
- 16
- 17
- 18

Answer Options

- 19
- 20
- 21

Select one or more options

[Clear All](#)

- 22
- 23
- 24

The analyst of Chennai Super Kings wants to predict the number of sixes Mahi will hit in his first match of the IPL 2020 based on his performance in the IPL till now.

- 28
- 29
- 30



- 31
- 32
- 33

The marketing team of upGrad wants to predict the number of new intakes in its Data Science Program for the next month.

- 34
- 35
- 36

- 37

A company wants to predict its expenditure based on the cost of raw materials.

Course 3

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8
- 9

The government of India wants to predict the rate of unemployment for the next quarter based on the data of this quarter.

176 min left

6. C3 Time Series Forecasting

[Previous](#)[Next](#)

Course 2

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

Answer Options

Select any one option

[Clear All](#) Simple moving average SARIMA Holt-Winters' SARIMAX

Course 3

1 2 3

4 5 6

7 8 9

10 11 12

Course 2

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30



31 32 33

34 35 36

37

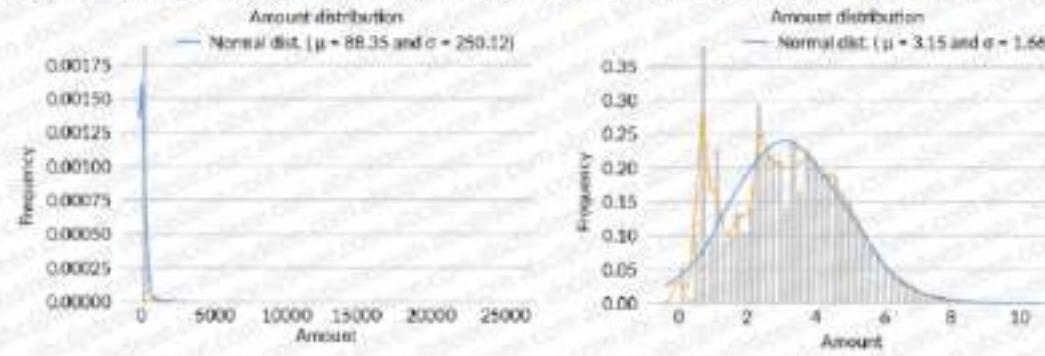
Course 3

1 2 3

4 5 6

7 8 9

Refer to the plot below which consists of two plots A and B. Plot A represents the non transformed column "Amount" while column B represents the transformed variable "Amount". Observe the plot B carefully and identify which kind of the transformation is applied to the variable "Amount"?



Answer Options

Select any one option

Clear Ans

 Standardisation Log Transformation Min-Max Scaling Power Transformation

Course 2

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

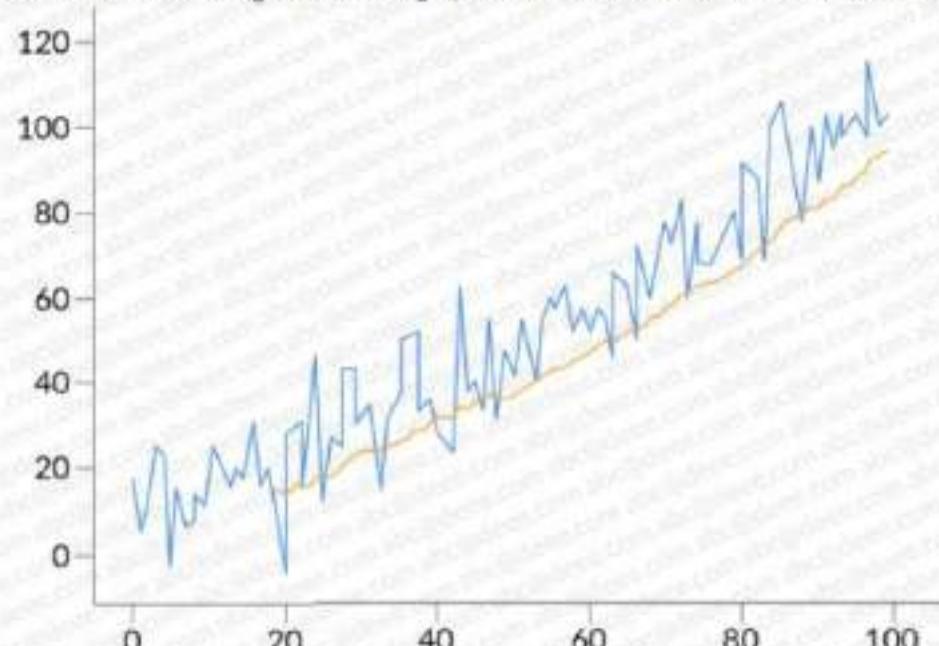
19 20 21

22 23 24

25 26 27

28 29 30

Consider the following time-series graph and choose the most correct statement:



<<

31 32 33

34 35 36

37

Answer Options

Select any one option

Clear Ans

 The time-series graph is stationary as mean and variance is constant. The time-series graph is non-stationary as mean is not constant. The time-series graph is non-stationary as mean and variance is increasing/changing.

Course 3

1 2 3

4 5 6

7 8 9

10 11 12

176 min left

3. Joins on MySQL

< Previous Next >

SQLITE

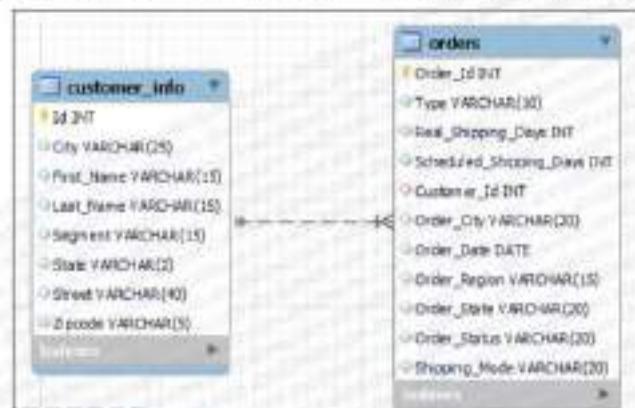
1 -- Enter your query here

Course 2

1 2 3
4 5 6
7 8 9
10 11 12
13 14 15
16 17 18
19 20 21
22 23 24
25 26 27
28 29 30
31 32 33
34 35 36
37

INPUT TABLE

You are given two tables - customer_info and orders whose schema is given below:

**QUERY:**Display all the details (all columns) from the **orders** table
where

- State is AZ and
- Street contains the word 'Silver'

Order them by **Order_Id** in ascending order.**Note** – The Id column in **customer_info** is common with the **Customer_Id** column in the **orders** table. You'll be required to use the LEFT JOIN to join the two tables.

Course 3

1 2 3
4 5 6
7 8 9
10 11 12
13 14 15
16 17 18
19 20 21
22 23 24
25 26 27
28 29 30
31 32 33
34 35 36
37

OUTPUT COLUMNS:All columns of the **orders** table.*Note that the coding console automatically converts the casing of the columns to upper case***PLEASE NOTE:**

- Use the original column names only. Any other aliases or column names would lead to an error.
- Keep the sequence of the columns the same as in the original table.

Schema

Table structure

department

Name	Type	Description
id	Int	
Name	varchar(25)	

177 min left

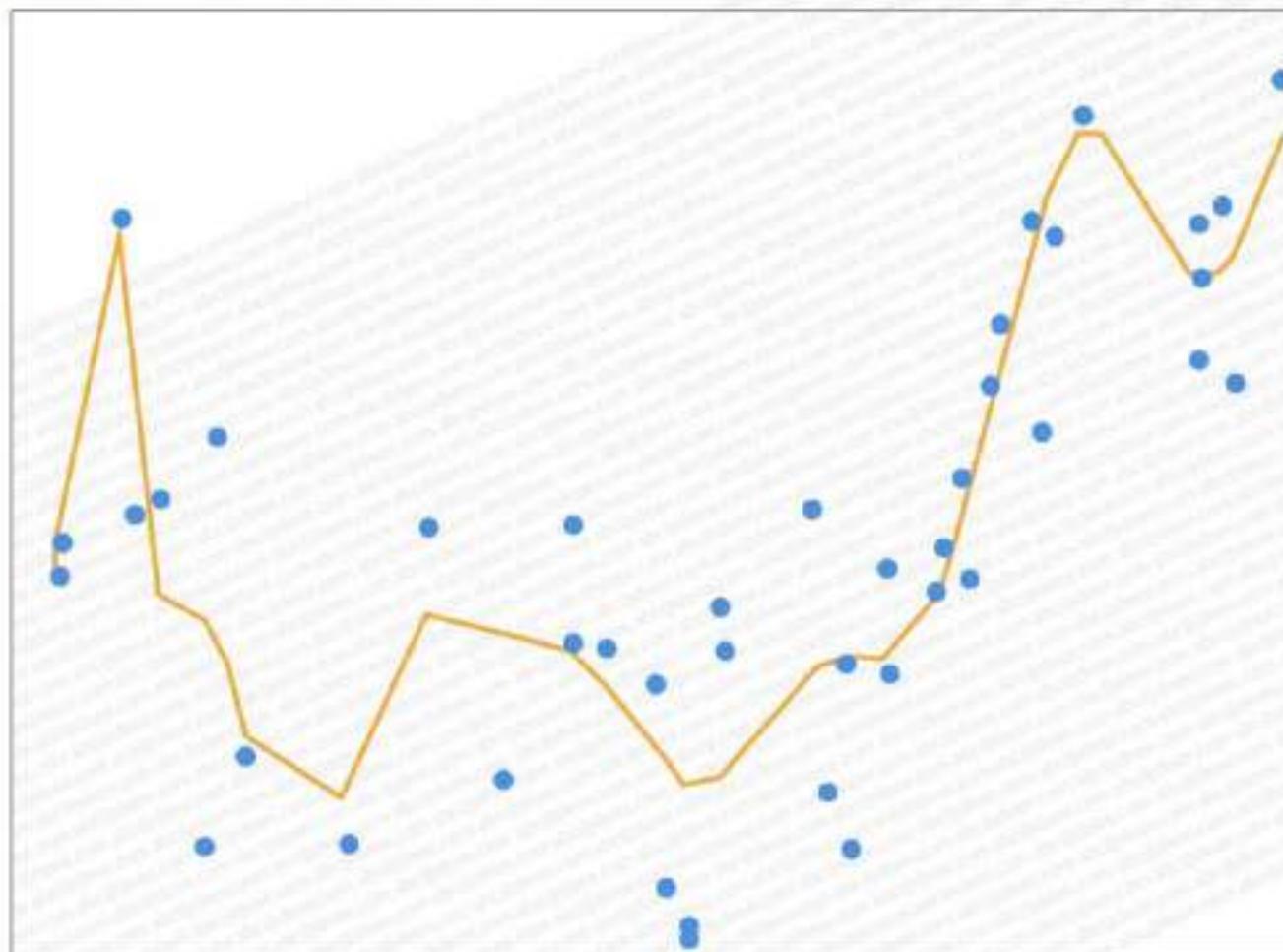
2. C3 Advanced ML

< Previous

Course 2

1	2	3
4	5	6
7	8	9
10	11	12
13	14	15
16	17	18
19	20	21
22	23	24
25	26	27
28	29	30
31	32	33
34	35	36
37		

Course 3



Answer Options

Select any one option

Clear

 He will get an underfitting model as the increase in the polynomial degree will not help reduce the model bias. He will get an overfitting model as the current model also overfits the train data.

22	23	24
25	26	27
28	29	30

31	32	33
----	----	----

177 min left

2. C3 Advanced ML

[◀ Previous](#)[Next ▶](#)

Course 2

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

31 32 33

34 35 36

37



Course 3

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

180 min left

2. C2 Multiple correct answer

[Previous](#)[Next](#)

Course 2

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

<<

28 29 30

31 32 33

34 35 36

37

Course 3

1 2 3

Which of the following is NOT a methodology by which you can identify the optimal number of clusters for K-means clustering? (More than one option may be correct)

Answer Options

Select one or more options

[Clear Ans](#) Dendrogram inspection method Elbow Method Single Linkage Method Silhouette score

Course 2

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8
- 9
- 10
- 11
- 12
- 13
- 14
- 15
- 16
- 17
- 18
- 19
- 20
- 21
- 22
- 23
- 24
- 25
- 26
- 27
- 28
- 29
- 30
- 31
- 32
- 33
- 34
- 35
- 36

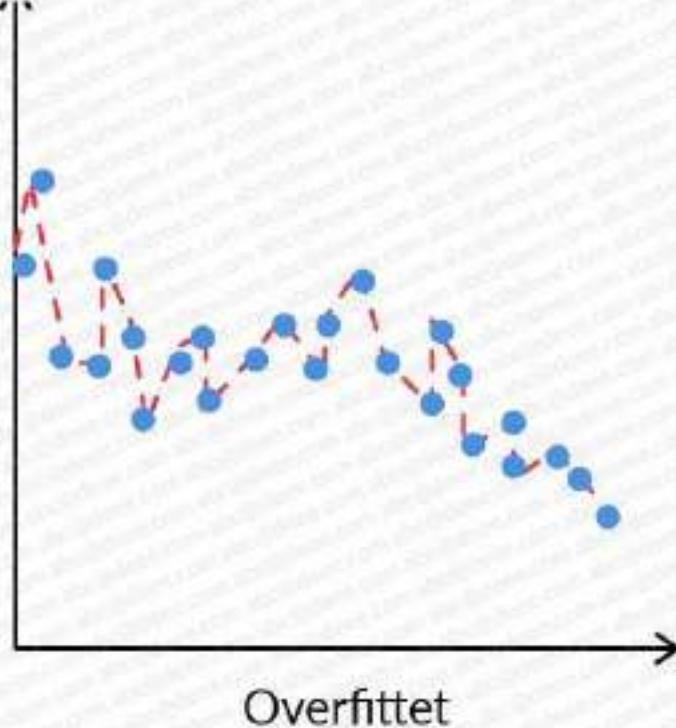
37

Course 3

Answer Options

Select any one option

Clear A

 If the model is using a decision tree regressor, then the depth of the tree needs to be increased. Increasing the number of datapoints in the training set We need to have more variables in the data.**Overfittet**

127 min left

37. kth Largest

< Previous Next >

Python 3

Course 2

Problem Statement

Given a list of integers and another integer 'k', find the **kth** largest integer in the list. If there are less than 'k' distinct elements in the list, then you need to output -1. (Refer to the sample inputs and outputs for details.)

Note: k-1 means you have to find the largest element

Input Format:

Line 1 contains a list of integers

Line 2 contains a positive integer $k > 0$

Output Format:

An integer, k th largest integer

Examples:**Sample Input 1:**

```
[2, 3, 1, 5, 6, 2, 1]
```

```
4
```

Sample Output 1:

```
2
```

Sample Input 2:

```
[2, 3, 1, 5, 6, 2, 1]
```

```
6
```

Sample Output 2:

```
-1
```

37

Function description

Course 3

Complete the **kthLargest** function in the editor below. It has the following parameter(s):

1 2 3
4 5 6
7 8 9
10 11 12
13 14 15
16 17 18

Name	Type	Description
k	INTEGER	
arr	INTEGER ARRAY	
Return	The function must return an INTEGER denoting the The k th largest integer in the array	

Constraints

```

1 import sys
2
3
4 def kthLargest(arr, k):
5     # Write your code here
6
7
8
9 def main():
10    import ast
11    arr = []
12    arr = ast.literal_eval(input())
13    k = int(input())
14    result = kthLargest(arr, k)
15    print(result)
16
17
18 if __name__ == "__main__":
19    main()

```

Console

Custom Test Case

127 min left

37. kth Largest

< Previous Next >

Python 3

Course 2

Problem Statement

Given a list of integers and another integer 'k', find the **kth** largest integer in the list. If there are less than 'k' distinct elements in the list, then you need to output -1. (Refer to the sample inputs and outputs for details.)

Note: k-1 means you have to find the largest element

Input Format:

Line 1 contains a list of integers

Line 2 contains a positive integer $k > 0$

Output Format:

An integer, k th largest integer

Examples:**Sample Input 1:**

```
[2, 3, 1, 5, 6, 2, 1]
```

```
4
```

Sample Output 1:

```
2
```

Sample Input 2:

```
[2, 3, 1, 5, 6, 2, 1]
```

```
6
```

Sample Output 2:

```
-1
```

37

Function description

Course 3

Complete the **kthLargest** function in the editor below. It has the following parameter(s):

1 2 3
4 5 6
7 8 9
10 11 12
13 14 15
16 17 18

Name	Type	Description
k	INTEGER	
arr	INTEGER ARRAY	
Return	The function must return an INTEGER denoting the The k th largest integer in the array	

Constraints

```

1 import sys
2
3
4 def kthLargest(arr, k):
5     # Write your code here
6
7
8
9 def main():
10    import ast
11    arr = []
12    arr = ast.literal_eval(input())
13    k = int(input())
14    result = kthLargest(arr, k)
15    print(result)
16
17
18 if __name__ == "__main__":
19    main()

```

Console

Custom Test Case

36. C2 Multiple correct answer

Course 2

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

<<

28 29 30

31 32 33

34 35 36

36

37

Course 3

1 2 3

ROC curve shows the tradeoff between the True Positive Rate (TPR) and the False Positive Rate (FPR). TPR and FPR are sensitivity and specificity respectively. The following function is written in Python using metrics package from the scikit-learn library for a ROC curve function.

```
def draw_roc(actual, probs):
    fpr tpr,thresholds=metrics.roc_curve(actual,probs,drop_intermediate=False)
    auc_score = metrics.roc_auc_score(actual, probs)
    return None
```

Which of the following statements are true? (More than one option may be correct.)

Answer Options

Select one or more options

Clear Ans

 The area under the ROC curve can be more than 1. The arguments passed in the above function are actual values of the target variable and the predicted values (i.e., 0 or 1) Larger the area under the curve, the better will be the model The arguments passed in the above function are actual values of the target variable and the respective predicted probabilities

35. C2 Clustering

Course 2

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

<<

28 29 30

31 32 33

34 35 36

37

Course 3

1 2 3

Silhouette metric for any i th point is given by: $S(i) = (b(i) - a(i)) / \max\{b(i), a(i)\}$

Which of the following is not true about Silhouette metric?

Answer Options

Select any one option

Clear Ans

 b(i) is the average distance from the nearest neighbour cluster(Separation). a(i) is the average distance from own cluster(Cohesion). If $S(i) = 1$ then the datapoint is similar to its own cluster. Silhouette Metric ranges from 0 to +1

34. C2-Basics of NLP and Text Mining

[Previous](#)[Next](#)

Course 2

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

Choose the correct option from the following.

The difference between '+' and '*' quantifier is ____.

Answer Options

Select any one option

[Clear Ans](#) '+' needs the preceding character to be present at least once whereas '*' does not need the same. '*' needs the character to be present at least once whereas '+' does not need the same. Both the quantifiers have the same functionality. None of the above

Course 3

1 2 3

177 min left

33. C2-Decision Trees

[Previous](#)[Next](#)

Course 2

- 1 2 3
- 4 5 6
- 7 8 9
- 10 11 12
- 13 14 15
- 16 17 18
- 19 20 21
- 22 23 24
- 25 26 27
- 28 29 30
- 31 32 **33**
- 34 35 36
- 37

Suppose you train a decision tree with the following data. Which feature should we split on at the root?

Answer Options

Select any one option

[Clear Ans](#) X Y Z Cannot be determined

Course 3

- 1 2 3

177 min left

32. C2-Basics of NLP and Text Mining

[Previous](#)[Next](#)

Course 2

- [1](#)
- [2](#)
- [3](#)
- [4](#)
- [5](#)
- [6](#)
- [7](#)
- [8](#)
- [9](#)
- [10](#)
- [11](#)
- [12](#)
- [13](#)
- [14](#)
- [15](#)
- [16](#)
- [17](#)
- [18](#)
- [19](#)
- [20](#)
- [21](#)
- [22](#)
- [23](#)
- [24](#)
- [25](#)
- [26](#)
- [27](#)
- [««](#)
- [32](#)
- [33](#)
- [34](#)
- [35](#)
- [36](#)
- [37](#)

Course 3

- [1](#)
- [2](#)
- [3](#)

Which of the following strings will match with the regular expression "^01*0\$"?

- 1. 0
- 2. 00
- 3. 0111111110

Answer Options

Select any one option

[Clear Ans](#)

- Only option 1
- Only option 3
- Both 1&2
- Both 2&3

177 min left

31. C2 linear Regression

[Previous](#)[Next](#)

Course 2

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

31 32 33

34 35 36

37

A scatterplot was plotted for two variables - age and income to find out how the income depends on the age of a person. It was found that as the income increases linearly with age, the variability in income also increases. This is a violation of which of the following assumptions of linear regression?

Answer Options

Select any one option

[Clear All](#) Homogeneity Heterogeneity Homoskedasticity Linearity

Course 3

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

177 min left.

31. C2 linear Regression

[Previous](#)[Next](#)

Course 2

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

31 32 33

34 35 36

37

Course 3

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

A scatterplot was plotted for two variables - age and income to find out how the income depends on the age of a person. It was found that as the income increases linearly with age, the variability in income also increases. This is a violation of which of the following assumptions of linear regression?

Answer Options

Select any one option:

Clear

 Homogeneity Heterogeneity Homoskedasticity Linearity Processing... Processing the answer...

No feedback

We cannot yet give you the status of this course for this question. We will update it once we have more information.

[View result](#)

180 min left

1. C2 Business Problem Solving

[Previous](#)[Next](#)

Course 2

1

2

3

4

5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

Answer Options

25

26

27



Select any one option

[Clear Ans](#)

28

29

30

 3>1>5>2>4>7>6>8

31

32

33

 3>2>1>5>4>7>6>8

34

35

36

37

 4>3>1>2>5>7>6>8

Course 3

1

2

3

 3>2>1>5>4>7>8>6

177 min left

30. C2 Logistic Regression

[Previous](#)[Next](#)

Course 2

- [1](#)
- [2](#)
- [3](#)
- [4](#)
- [5](#)
- [6](#)
- [7](#)
- [8](#)
- [9](#)
- [10](#)
- [11](#)
- [12](#)
- [13](#)
- [14](#)
- [15](#)
- [16](#)
- [17](#)
- [18](#)
- [19](#)
- [20](#)
- [21](#)

Recall the telecom churn example, if the log odds for churn are equal to 0 for a customer, then that means -

- [22](#)
- [23](#)
- [24](#)
- [25](#)
- [26](#)
- [27](#)
- [28](#)
- [29](#)
- [30](#)
- [31](#)
- [32](#)
- [33](#)
- [34](#)
- [35](#)
- [36](#)
- [37](#)

Answer Options

Select any one option

[Clear All](#)

- There is no chance of the customer churning
- The probability of the customer churning is equal to the probability of the customer not churning
- The probability of the customer churning is very small compared to the probability of the customer not churning
- The probability of the customer churning is very large compared to the probability of the customer not churning

Course 3

- [1](#)
- [2](#)
- [3](#)
- [4](#)
- [5](#)
- [6](#)
- [7](#)
- [8](#)
- [9](#)
- [10](#)
- [11](#)
- [12](#)
- [13](#)
- [14](#)
- [15](#)
- [16](#)
- [17](#)
- [18](#)
- [19](#)
- [20](#)
- [21](#)
- [22](#)
- [23](#)
- [24](#)

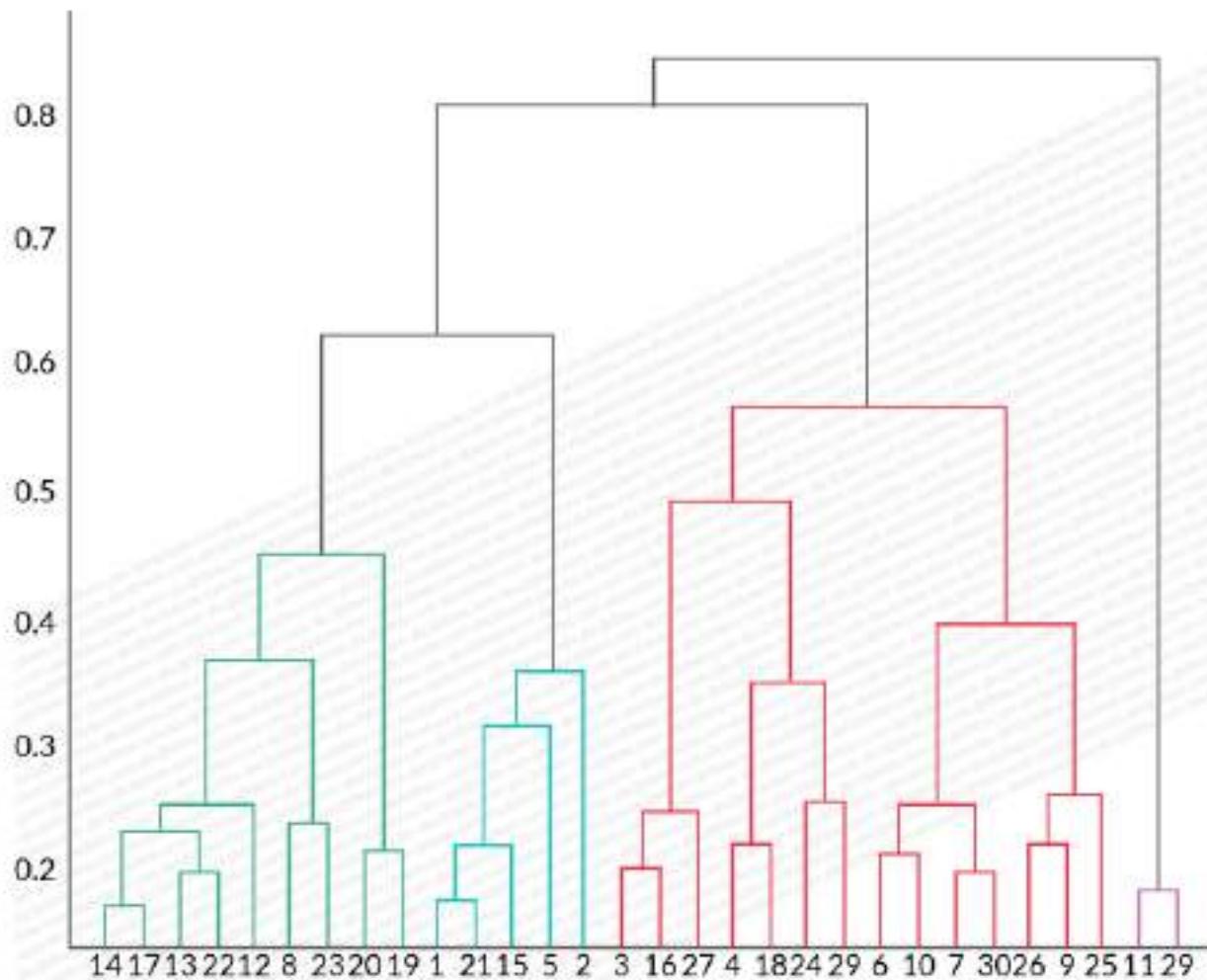
Dataset 2

1	1	3
4	5	6
7	8	9
10	11	12
13	14	15
16	17	18
19	20	21
22	23	24
25	26	27
28	29	30
31	32	33
34	35	36
37		

Dataset 1

1	2	3
4	5	6
7	8	9
10	11	12
13	14	15
16	17	18
19	20	21
22	23	24
25	26	27
28	29	30
31	32	33
34	35	36
37		

You obtained the following dendrogram after performing K-means clustering on a dataset. Which of the following statements can be concluded from the dendrogram?



Answer Options

Select any one option

- The initial number of clusters is 6
- There are 25 data points used in the above clustering algorithm
- Single linkage is used to define the distance between two clusters in the above dendrogram

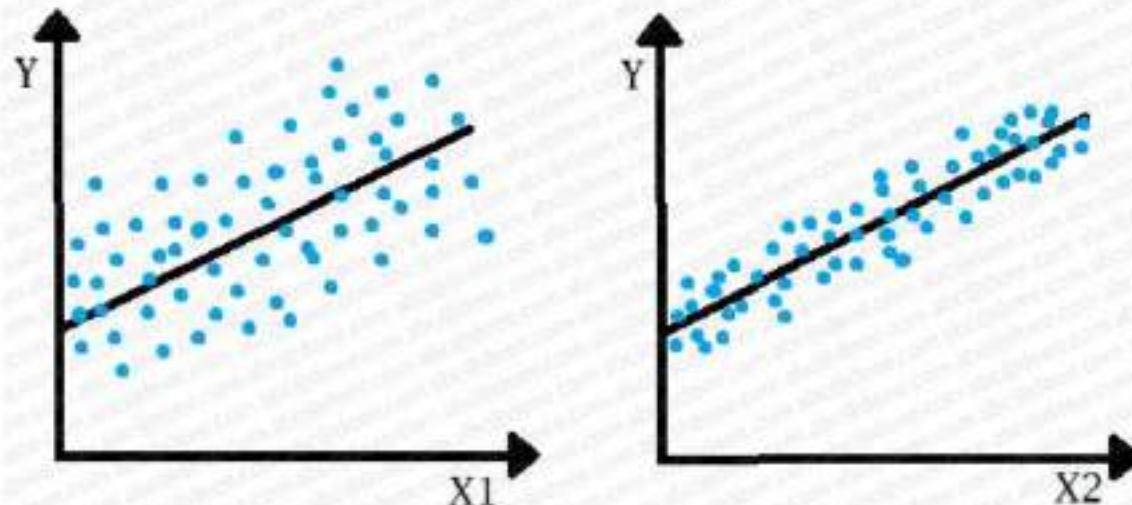
The above dendrogram is not suitable for K-Means clustering.

28. C2 linear Regression

Course 2

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8
- 9
- 10
- 11
- 12
- 13
- 14
- 15
- 16
- 17
- 18
- 19
- 20
- 21
- 22
- 23
- 24
- 25
- 26
- 27
- 28
- 29
- 30
- 31
- 32
- 33
- 34
- 35
- 36

For the same dependent variable Y, two models were created using the independent variables X1 and X2. The following two graphs represent the fitted line on the scatterplot. (Both of the graphs are on the same scale)



Which of the following is true about the residuals in these two models?

Answer Options

Select any one option

Clear A

Course 3

The sum of residuals in model 2 is higher than model 1

The sum of residuals in model 1 is higher than model 2

?

Both have the same sum of residuals

Nothing can be said about the sum of residuals from the graph

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8
- 9
- 10
- 11
- 12
- 13
- 14
- 15
- 16
- 17
- 18

178 min left

28. C2 linear Regression

[Previous](#)[Next](#)

Course 2

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30



31 32 33

34 35 36

37

Course 3

1 2 3

4 5 6

7 8 9

For the same dependent variable Y, two models were created using the independent variables X1 and X2. The following two graphs represent the fitted on the scatterplot. (Both of the graphs are on the same scale)

Which of the following is true about the residuals in these two models?

Answer Options

Select any one option

[Clear All](#) The sum of residuals in model 2 is higher than model 1 The sum of residuals in model 1 is higher than model 2 Both have the same sum of residuals Nothing can be said about the sum of residuals from the graph

178 min left

27. C2-Basics of NLP and Text Mining

< Previous

Next

Course 2

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30



31 32 33

34 35 36

37

Answer Options

Select any one option

Clear Ans

3

4

5

6

Course 3

1 2 3

4 5 6

7 8 9

10 11 12

178 min left

26. C2 Multiple correct answer

[Previous](#)[Next](#)

Course 2

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30



31 32 33

34 35 36

37

Course 3

1 2 3

4 5 6

7 8 9

Which of the following command correctly builds a logistic regression model in Python? (More than one option may be correct.)

Answer Options

Select one or more options

[Clear All](#) from sklearn.linear_model import LogisticRegression

lr = LogisticRegression()

lr.fit(X_train, y_train)

 import statsmodel.api as sm

lr = sm.GLM(y_train,(sm.add_constant(X_train)),

family = sm.families.Binomial())

lr.fit()

 from sklearn.linear_model import LogisticRegression

lr = LogisticRegression()

lr.predict(X_train, y_train)

 import statsmodel.api as sm

178 min left

25. C2-Decision Trees

[Previous](#)[Next](#)

Course 2

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8
- 9
- 10
- 11
- 12
- 13
- 14
- 15
- 16
- 17
- 18
- 19
- 20
- 21
- 22
- 23
- 24
- 25
- 26
- 27
- 28
- 29
- 30
- 31
- 32
- 33
- 34
- 35
- 36
- 37

Which of the following metrics measures how often a randomly chosen element would be incorrectly identified?

Answer Options

Select any one option

[Clear All](#) Entropy Information Gain Gini Index

Course 3

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8
- 9

 None of these

178 min left

25. C2-Decision Trees

[Previous](#)[Next](#)

Course 2

- | | | |
|----|----|----|
| 1 | 2 | 3 |
| 4 | 5 | 6 |
| 7 | 8 | 9 |
| 10 | 11 | 12 |
| 13 | 14 | 15 |
| 16 | 17 | 18 |
| 19 | 20 | 21 |
| 22 | 23 | 24 |
| 25 | 26 | 27 |
| 28 | 29 | 30 |
| 31 | 32 | 33 |
| 34 | 35 | 36 |
| 37 | | |

Which of the following metrics measures how often a randomly chosen element would be incorrectly identified?

Answer Options

Select any one option

[Clear Ans](#) Entropy Information Gain Gini Index None of these

Course 3

- | | | |
|----|----|----|
| 1 | 2 | 3 |
| 4 | 5 | 6 |
| 7 | 8 | 9 |
| 10 | 11 | 12 |

178 min left

24. C2 linear Regression

[Previous](#)[Next](#)

Course 2

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30



31 32 33

34 35 36

37

Course 3

1 2 3

4 5 6

7 8 9

Which of the following is true regarding the error terms in linear regression?

Answer Options

Select any one option

[Clear All](#) The sum of residuals should be zero The sum of residuals should be lesser than zero The sum of residuals should be greater than zero There is no such restriction on what the sum of residuals should be

178 min left

23. C2 Logistic Regression

[Previous](#)[Next](#)

Course 2

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

31 32 33

34 35 36

37

Given an imbalanced dataset, the ratio of positive to negative class is 1:10000. You run a logistic regression model and find out that the model has a high value of precision and a low value of recall. Which of the following statements is true?

Answer Options

Select any one option

[Clear All](#) The class is handled well by the data The model is not able to detect the class, but when it does it is highly trustable The model is able to detect the class but it includes data points from the other class as well

1 2 3

4 5 6

7 8 9

10 11 12

Course 3

 The class is handled poorly by the data

178 min left

22. C2 Logistic Regression

[◀ Previous](#)[Next ▶](#)

Course 2

1 2 3
4 5 6
7 8 9
10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

31 32 33

34 35 36

37

Course 3

1 2 3

4 5 6

7 8 9

Consider the following confusion matrix.

Total=500	Actual Positive	Actual Negative
Predicted Positive	196	20
Predicted Negative	28	256

Which among the following is the lowest for the given confusion matrix?

Answer Options

Select any one option

[Clear Ans](#)

Accuracy

Precision

Sensitivity

Specificity

178 min left

21. C2 linear Regression

[Previous](#)[Next](#)

Course 2

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

How is regression different from classification?

Answer Options

Select any one option

[Clear Ans](#) One is supervised while the other is unsupervised One is iterative while the other is closed form In regression, the response variable is numeric while it is categorical in classification

Course 3

1 2 3

4 5 6

7 8 9

 None of the above

178 min left

20. C2 Clustering

[Previous](#)[Next](#)

Course 2

Which of the following statements is NOT true?

- 1 2 3
- 4 5 6
- 7 8 9
- 10 11 12
- 13 14 15
- 16 17 18

Answer Options

- 19 20 21
- 22 23 24
- 25 26 27
- 28 29 30
- 31 32 33
- 34 35 36
- 37

Select any one option

[Clear Ans](#)

- Each time the clusters are made during the K-means algorithm, the centroid is updated.
- The cluster centres that are computed in the K-means algorithm are given by centroid value of the cluster points.
- Standardization of the data is not important before applying Euclidean distance as a measure of similarity/dissimilarity.

Course 3

- 1 2 3
- 4 5 6
- 7 8 9
- 10 11 12

- The centroid of a column with data points 25, 32, 34 and 23 is 28.5

176 min left

19. C2 Multiple correct answer

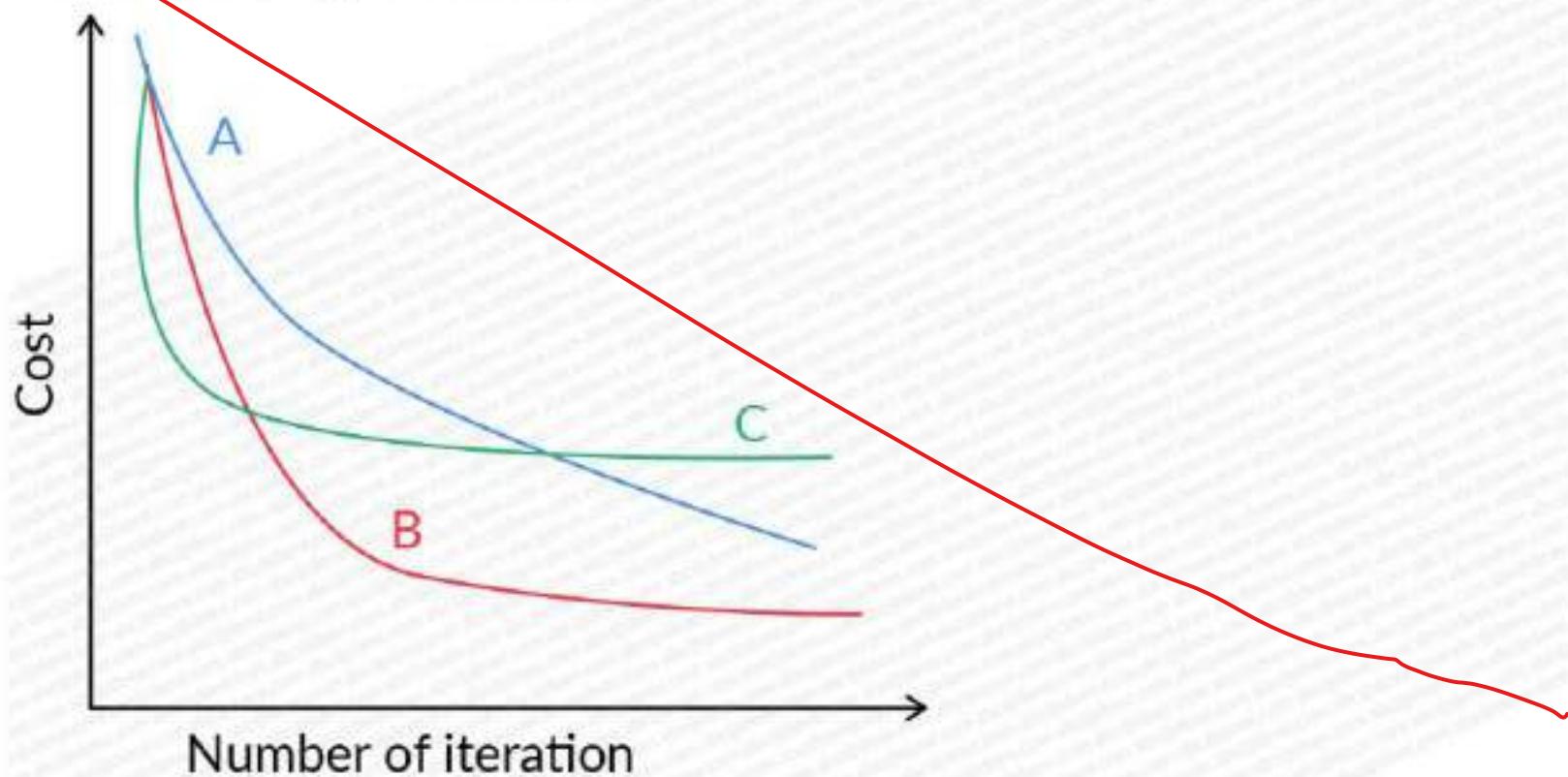
< Previous

Course 2

- | | | |
|----|----|----|
| 1 | 2 | 3 |
| 4 | 5 | 6 |
| 7 | 8 | 9 |
| 10 | 11 | 12 |
| 13 | 14 | 15 |
| 16 | 17 | 18 |
| 19 | 20 | 21 |
| 22 | 23 | 24 |
| 25 | 26 | 27 |
| 28 | 29 | 30 |
| 31 | 32 | 33 |
| 34 | 35 | 36 |
| 37 | | |

Course 3

Observe the following cost function graph with different learning rates.



Which of the following statements are true about the learning rate? (More than one option may be correct.)

Answer Options:

Select one or more options

Clear

 The learning rate of curve C is highest among all curves The learning rate for curve B is lower than A The learning rate for curve B is higher than A

176 min left

19. C2 Multiple correct answer

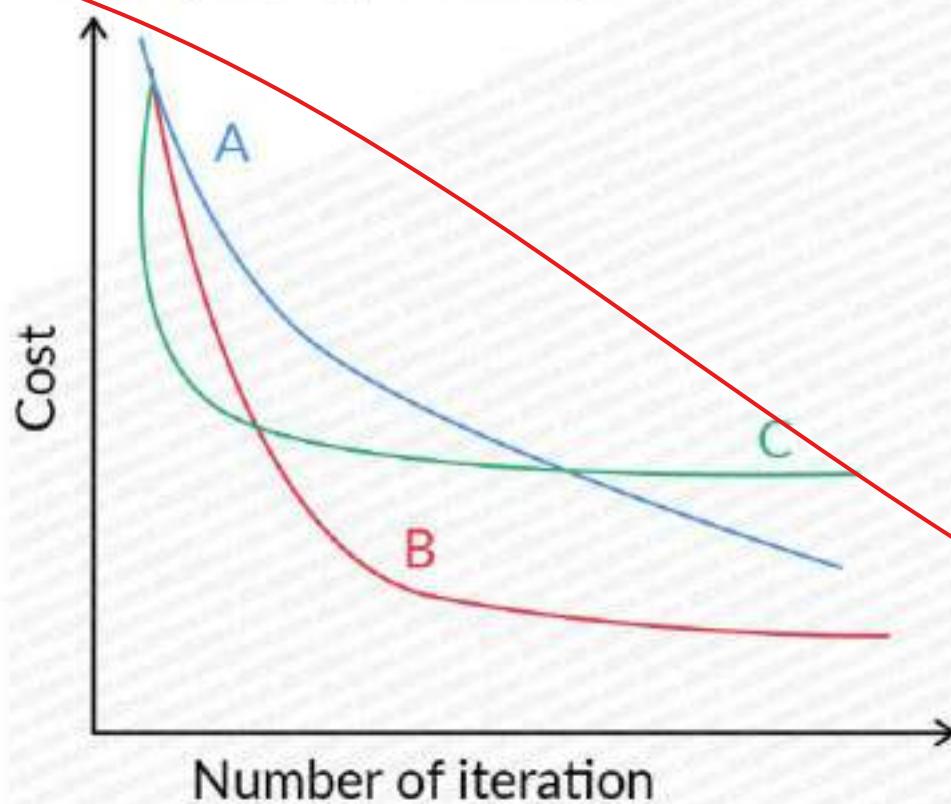
< Previous

Course 2

- | | | |
|----|----|----|
| 1 | 2 | 3 |
| 4 | 5 | 6 |
| 7 | 8 | 9 |
| 10 | 11 | 12 |
| 13 | 14 | 15 |
| 16 | 17 | 18 |
| 19 | 20 | 21 |
| 22 | 23 | 24 |
| 25 | 26 | 27 |
| 28 | 29 | 30 |
| 31 | 32 | 33 |
| 34 | 35 | 36 |
| 37 | | |

Course 3

Observe the following cost function graph with different learning rates.



Which of the following statements are true about the learning rate? (More than one option may be correct.)

Answer Options:

Select one or more options.

Clear

 The learning rate of curve C is highest among all curves The learning rate for curve B is lower than A The learning rate for curve B is higher than A

179 min left

19. C2 Multiple correct answer

[Previous](#)[Next](#)

Course 2

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

<<

28 29 30

31 32 33

34 35 36

37

Course 3

1 2 3

Observe the following cost function graph with different learning rates.

Which of the following statements are true about the learning rate? (More than one option may be correct.)

Answer Options

Select one or more options

[Clear Ans](#) The learning rate of curve C is highest among all curves The learning rate for curve B is lower than A The learning rate for curve B is higher than A The learning rate of curve C is the smallest among all curves None of the above

179 min left

18. C2 Logistic Regression

[Previous](#)[Next](#)

Course 2

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27



28 29 30

31 32 33

34 35 36

37

Course 3

1 2 3

You have built a Logistic Regression model that is trying to predict whether a loan is approved or not based on a person's FICO score. Here are the model parameters: Intercept (B_0) = -9.346 and coefficient of FICO score = 0.0146. Given these parameters, can you calculate the probability of a loan getting approved for someone with a FICO score of 655?

Answer Options

Select any one option

[Clear Ans](#) 0.35 0.45 0.55 0.65

179 min left

17. C2 linear Regression

[Previous](#)[Next](#)

Course 2

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27



28 29 30

31 32 33

34 35 36

37

Course 3

1 2 3

Suppose you run a regression with one of the feature variables T, with all the remaining feature variables. The R-squared of this model was found out to be 0.8. What will be the VIF for the variable T?

Answer Options

Select any one option

[Clear Ans](#) 1.56 2.77 3.33 5.00

16. C2 Multiple correct answer

[Previous](#)[Next](#)

Course 2

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

Answer Options

Select one or more options

[Clear Ans](#) The null hypothesis for a simple linear regression model is $H_0: \beta_1 = 0$ If the p-value turns out to be greater than 0.05 for β_1 , it means β_1 is significant If β_1 turns out to be insignificant, that means there is no relationship between the dependent and independent variable

37

Course 3

1 2 3

 If the p-value turns out to be less than 0.05 for β_0 , it means that β_0 is non-zero

179 min left

15. C2 linear Regression

[Previous](#)[Next](#)

Course 2

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

In linear regression, the metric F-statistic is used to determine

Answer Options

Select any one option

[Clear Ans](#) the significance of the individual beta coefficients the variance explanation strength of the model the significance of the overall model fit Both A & C

37

Course 3

1 2 3

179 min left

14. C2 Logistic Regression

[Previous](#)[Next](#)

Course 2

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

Which of the following is correct for a logistic regression model?

Answer Options

Select any one option

[Clear Ans](#) The independent variables should not be multicollinear. The dependent variable should follow Normal Distribution. The log odds in a logistic regression model lies between 0 and 1. F1-score is always the best metric for evaluating a logistic regression model.

Course 3

1 2 3

13. C2 Business Problem Solving

Course 2

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

Select any one option

[Clear Ans](#) Statement 1 is correct and Statement 2 is wrong Statement 2 is correct and Statement 1 is wrong Both the statements are correct None of the statements are correct

Course 3

1 2 3

179 min left

12. C2 Clustering

[Previous](#)[Next](#)

Course 2

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

Which of the following is not true for Hopkins statistic?

Answer Options

Select any one option

[Clear Ans](#) Hopkins statistic decides if the data is suitable for clustering or not Hopkins statistic lie between -1 and 1 If the Hopkins statistic comes out to be 0, then the data is uniformly distributed If the Hopkins statistic comes out to be 1, then the data highly suitable for clustering

Course 3

1 2 3

179 min left

11. C2 Logistic Regression

[Previous](#)[Next](#)

Course 2

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8
- 9
- 10
- 11
- 12
- 13
- 14
- 15
- 16
- 17
- 18
- 19
- 20
- 21
- 22
- 23
- 24
- 25
- 26
- 27

Consider the following univariate logistic model:

$$Y = \beta_0 + \beta_1 X_1$$

Which of the following statements is NOT true?

Answer Options

Select any one option

[Clear Ans](#)

The maximum likelihood estimation determines the best combination of β_0 and β_1 .

If β_1 is increased by 1 unit, Y increases by 1 unit.

β_0 is the y-intercept

If β_1 is increased by 1 unit, log-odds increases by 1 unit.

Course 3

- 1
- 2
- 3

179 min left

11. C2 Logistic Regression

[Previous](#)[Next](#)

Course 2

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8
- 9
- 10
- 11
- 12
- 13
- 14
- 15
- 16
- 17
- 18
- 19
- 20
- 21
- 22
- 23
- 24
- 25
- 26
- 27
- [«](#)
- 28
- 29
- 30
- 31
- 32
- 33
- 34
- 35
- 36
- 37

Consider the following univariate logistic model:

$$Y = \beta_0 + \beta_1 X_1$$

Which of the following statements is NOT true?

Answer Options

Select any one option

[Clear Ans](#)

- The maximum likelihood estimation determines the best combination of β_0 and β_1 .
- If β_1 is increased by 1 unit, Y increases by 1 unit.
- β_0 is the y-intercept
- If β_1 is increased by 1 unit, log-odds increases by 1 unit.

10. C2 Business Problem Solving

Course 2

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27



Logistic regression

28 29 30

31 32 33

34 35 36

37

Course 3

1 2 3

The coronavirus disease (COVID-19), was declared a pandemic by the World Health Organization (WHO) in February 2020. Currently, there are no vaccines or treatments that have been officially approved by WHO after clinical trials. India has not seen the peak of infection yet and the number of infections is touching a new height daily. The business unit of an Indian health and hygiene company approaches you and know "Why the sales of masks is decreasing despite the number of corona infections increasing daily". Answer the following questions:

Suppose you mapped the above problem statement with a classification problem, either a customer will buy a mask or not. Your will build _____ model as your initial solution.

Answer Options

Select any one option

[Clear Ans](#) Neural Network Logistic regression Decision tree All of the above

179 min left

9. C2 Clustering

[Previous](#)[Next](#)

Course 2

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27



28 29 30

31 32 33

34 35 36

37

Initialising the following command in Python will result in the following: `model_clus = KMeans(n_clusters = 6, max_iter=50)`

Answer Options

Select any one option

[Clear Ans](#) Run maximum 6 iterations Run maximum 40 iterations Create 6 final clusters Create 50 final clusters

Course 3

1 2 3

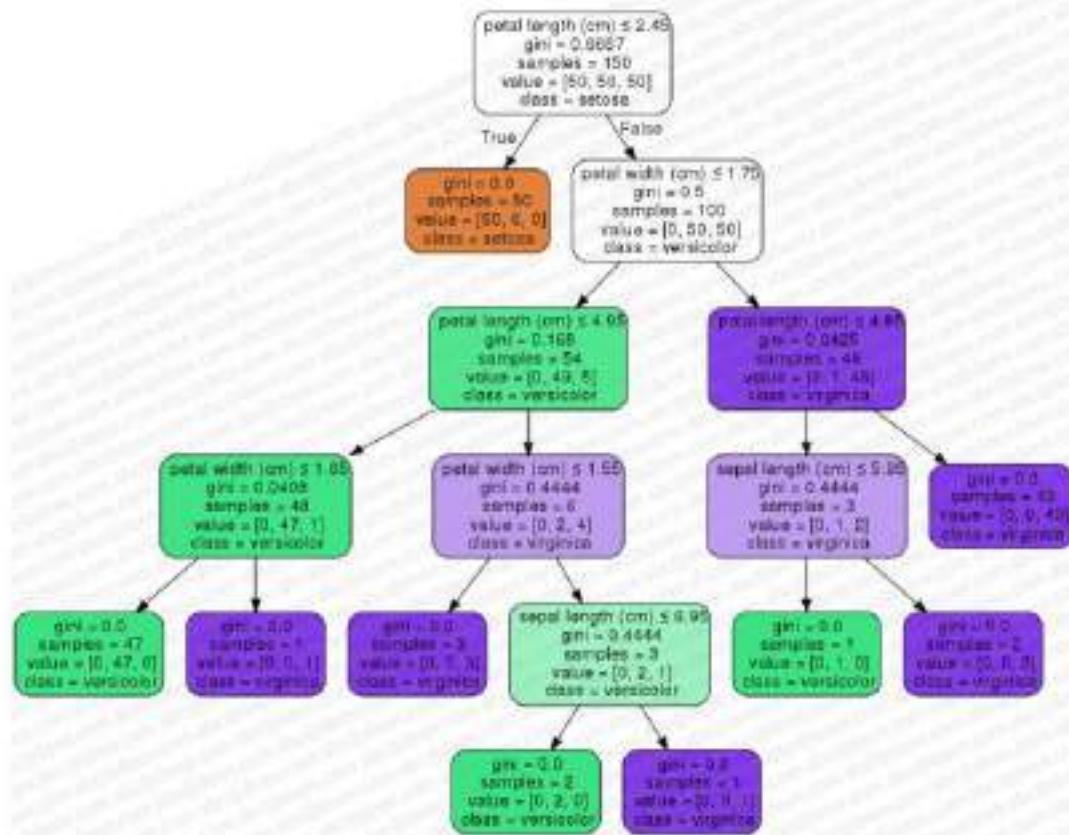
Course 2

- 1 2 3
 4 5 6
 7 8 9
 10 11 12
 13 14 15
 16 17 18
 19 20 21
 22 23 24
 25 26 27
 28 29 30
 31 32 33
 34 35 36

Course 3

- 1 2 3
 4 5 6
 7 8 9
 10 11 12
 13 14 15
 16 17 18
 19 20 21
 22 23 24
 25 26 27
 28 29 30
 31 32 33

Refer to the decision tree given below and choose the statement that is correct as per this tree.



Answer Options

Select any one option

Clear

The tree given above will show very good performance on the train data.

The tree given above is an underfitting tree.

If the petal length is more than 2.45, then it is equally likely that the flower is either setosa or virginica.

179 min left

8. C2-Decision Trees

[Previous](#)[Next](#)

Course 2

- 1 2 3
- 4 5 6
- 7 8 9
- 10 11 12
- 13 14 15
- 16 17 18
- 19 20 21
- 22 23 24
- 25 26 27
- 28 29 30
- 31 32 33
- 34 35 36
- 37

Refer to the decision tree given below and choose the statement that is correct as per this tree.

Answer Options

Select any one option

[Clear Ans](#)

- The tree given above will show very good performance on the train data.
- The tree given above is an underfitting tree.
- If the petal length is more than 2.45, then it is equally likely that the flower is either setosa or virginica.
- Both B and C

Course 3

- 1 2 3

179 min left

7. GroupBy and OrderBy

[Previous](#)[Next](#)

SQLITE

Course 2

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

31 32 33

34 35 36

37

Course 3

1 2 3

INPUT TABLE

You're given a **orders** table and the columns in the orders table are shown below:

orders	
Order_Id	INT
Type	VARCHAR(10)
Real_Shipping_Days	INT
Scheduled_Shipping_Days	INT
Customer_Id	INT
Order_City	VARCHAR(20)
Order_Date	DATE
Order_Region	VARCHAR(15)
Order_State	VARCHAR(20)
Order_Status	VARCHAR(20)
Shipping_Mode	VARCHAR(20)
Indexes	

QUERY

- Calculate count of all the orders
 - **where** the *Order_State* is **Gujarat**
 - **where** the *Order_Status* is **PENDING**.
 - **Note** - Use the alias of **oc** for count of orders.
- **Group the results by** *Order_City*
- **Order them by** *oc & Order_City* in **ascending order**.

OUTPUT COLUMNS

oc, Order_City

Console

Custom Test Case

177 min left

23. C3 Advanced ML

[Previous](#)[Next](#)

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27



28 29 30



31 32 33



34 35 36



37 38 39



40 41 42



43 44 45

Suppose an imbalanced data set has a class ratio of 1:5, and you want to run a cross-validation scheme to evaluate a model's performance. If you apply a stratified k-fold to generate the train-test folds, what will be the distribution of the classes in the test split?

Answer Options

Select any one option

[Clear Ans](#) 1:5 2:3 1:6 None of these

177 min left

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

31 32 33

34 35 36

37 38 39

40 41 42

43 44 45

22. C2 Business Problem Solving

[Previous](#)[Next](#)

The coronavirus disease (COVID-19), was declared a pandemic by the World Health Organization (WHO) in February 2020. Currently, there are no vaccines or treatments that have been officially approved by WHO after clinical trials. India has not seen the peak of infection yet and the number of infections is touching a new height daily. The business unit of an Indian health and hygiene company approaches you and know "Why the sales of masks is decreasing despite the number of corona infections increasing daily". Answer the following questions:

Suppose you mapped the above problem statement with a classification problem, either a customer will buy a mask or not. You will build _____ model as your initial solution.

Answer Options

Select any one option

[Clear Ans](#) Neural Network Logistic regression Decision tree All of the above

178 min left

21. C2 Logistic Regression

[Previous](#)[Next](#)

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

31 32 33

34 35 36

37 38 39

40 41 42

43 44 45

The output of a logistic model is:

Answer Options

Select any one option

[Clear Ans](#) 0 or 1 Any value between 0 and 1 0.5 Depends on the business problem

178 min left

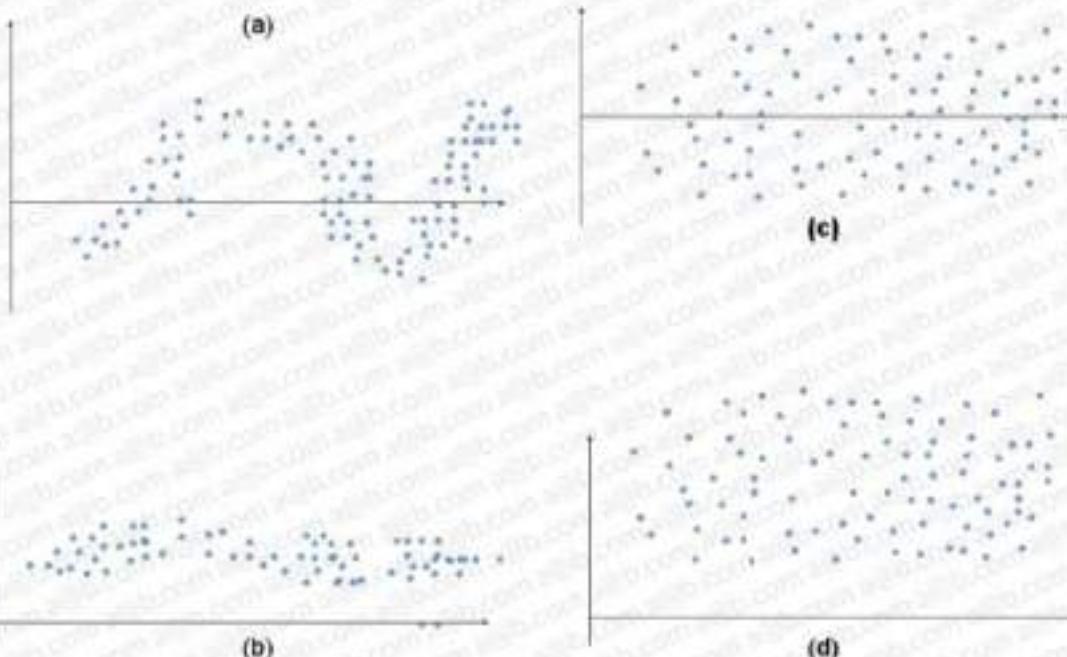
20. C2 linear Regression

[Previous](#)[Next](#)

MCQs

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8
- 9
- 10
- 11
- 12
- 13
- 14
- 15
- 16
- 17
- 18
- 19
- 20
- 21
- 22
- 23
- 24
- 25
- 26
- 27
- 28
- 29
- 30
- 31
- 32
- 33
- 34
- 35
- 36
- 37
- 38
- 39
- 40
- 41
- 42
- 43
- 44
- 45

The distribution of error terms in a linear regression model should look like (the horizontal line represents $y=0$):



Answer Options

Select any one option

[Clear Ans](#)

- A
- B
- C

178 min left

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

19. C3 Time Series Forecasting

[Previous](#)[Next](#)

You are an analyst at a retail company in India. Owing to the COVID-19 pandemic, there is a huge demand for sanitisers, and the rate of infection is following an upward trajectory (a higher number of infections daily). You have access to past 30 days data for the demand of sanitisers. Which of the following models would work best for forecasting the demand for sanitisers for the next 5 days with highest accuracy?

Answer Options

Select any one option

[Clear Ans](#) Seasonal Auto-Regressive Integrated Moving Average Model Simple Average Forecast Method Simple Exponential Smoothing Method Holt-Winters' Method

178 min left

18. C2 Logistic Regression

[Previous](#)[Next](#)

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

Which of the following statements is NOT true?

Answer Options

Select any one option

[Clear Ans](#) In the case of a fair coin, the odds of getting heads is 1 The error values of linear and logistic regression have to be normally distributed Specificity decreases with an increase in sensitivity As TPR increases, FPR also increases

<<

37 38 39

40 41 42

43 44 45

178 min left

17. C3 Time Series Forecasting

[Previous](#)[Next](#)

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

<<

28 29 30

31 32 33

34 35 36

37 38 39

40 41 42

43 44 45

The business administration of a five-year-old online retail company has observed that the revenue touches a new record once every quarter after the appraisal for the sales employee is released. The company wants to predict its revenue for the next quarter. You are working as an analytics consultant and the company approached you with its monthly and other relevant data to help them out. Which of the following methods of forecasting will you use for the best prediction/accuracy?

Answer Options

Select any one option

[Clear Ans](#) Simple moving average SARIMA Holt-Winters' SARIMAX

178 min left

16. C2 linear Regression

[Previous](#)[Next](#)

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

 A, C, and D

28 29 30

31 32 33

34 35 36

37 38 39

40 41 42

43 44 45

Which of the following assumptions do we make while building a simple linear regression model? (assume X and y to be independent and dependent variables respectively).

- A) There is a linear relationship between X and y.
- B) X and y are normally distributed.
- C) Error terms are independent of each other.
- D) Error terms have constant variance.

Answer Options

Select any one option

[Clear Ans](#) A, B, C and D A, C, and D A, B and C B, C and D

178 min left

15. C2 Clustering

[Previous](#)[Next](#)

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

<<

28 29 30

31 32 33

34 35 36

37 38 39

40 41 42

43 44 45

Silhouette metric for any i th point is given by: $S(i) = (b(i) - a(i)) / \max\{b(i), a(i)\}$

Which of the following is not true about Silhouette metric?

Answer Options

Select any one option

[Clear Ans](#) b(i) is the average distance from the nearest neighbour cluster(Separation). a(i) is the average distance from own cluster(Cohesion). If $S(i) = 1$ then the datapoint is similar to its own cluster. Silhouette Metric ranges from 0 to +1

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

Using the Gini Index as the impurity, find the root node to construct a decision tree from the information given below.

A	B	Class Label
T	F	+
T	T	+
T	T	+
T	F	-
T	T	+
F	F	-
F	F	-
F	F	-
T	T	-
T	F	-

Answer Options

<<

Select any one option

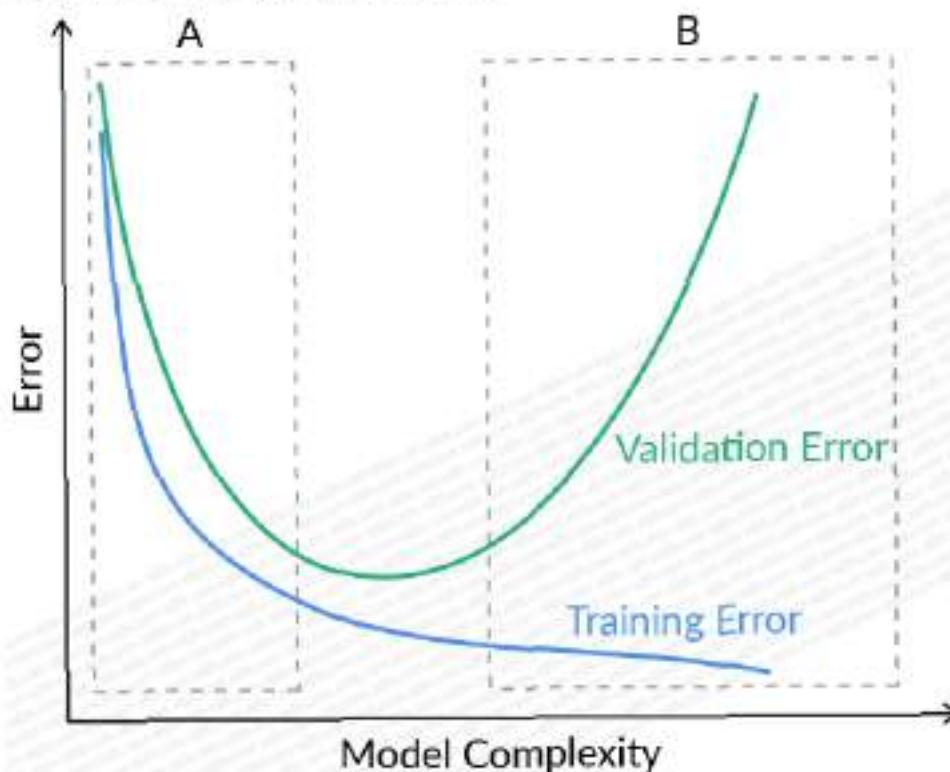
Clear Ans

 A B Both A and B have the same Gini Index value and so, we cannot make the split. Information provided is insufficient

MCQs

1	2	3
4	5	6
7	8	9
10	11	12
13	14	15
16	17	18
19	20	21
22	23	24
25	26	27
28	29	30
31	32	33
34	35	36
37	38	39
40	41	42
43	44	45
46	47	48
49	50	51
52	53	54
55	56	57
58	59	60
61	62	63
64	65	66
67	68	69

Refer to the image and choose the best option that represents loc A and B.



Answer Options:

Select one or more options:

 A- Underfitting, B- Good Model A- Good Model, B- Overfitting ✓ A- Underfitting, B- Overfitting A- Overfitting, B- Underfitting

178 min left

12. C2 Logistic Regression

< Previous

Next

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30



31 32 33

34 35 36

37 38 39

40 41 42

43 44 45

46 47 48

49 50 51

52 53 54

Consider the following two statements:

Statement 1: Suppose the value of Precision and Recall for a model are 0.65 and 0.75 respectively. Then the value of F1-score will be ~0.696.**Statement 2:** Mean squared error is a metric that can be used to evaluate a logistic regression model.

Answer Options

Select any one option

Clear Ans

 Statement 1 is wrong and statement 2 is correct Statement 1 is correct and statement 2 is wrong Both the statements are correct None of the statements are correct

179 min left

11. C2 Clustering

[Previous](#)[Next](#)

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

Which of the following statements is NOT true?

19 20 21

Answer Options

Select any one option

[Clear Ans](#) The cluster centers that are computed in the K-means algorithm are given by the centroid value of the cluster points. Standardization of the data is important before applying Euclidean distance as a measure of similarity/dissimilarity The centroid of a column with data points 25, 32, 34 and 23 is 28.5 The Euclidean distance between two points (10,2) and (4,5) is 7.

31 32 33

34 35 36

37 38 39

40 41 42

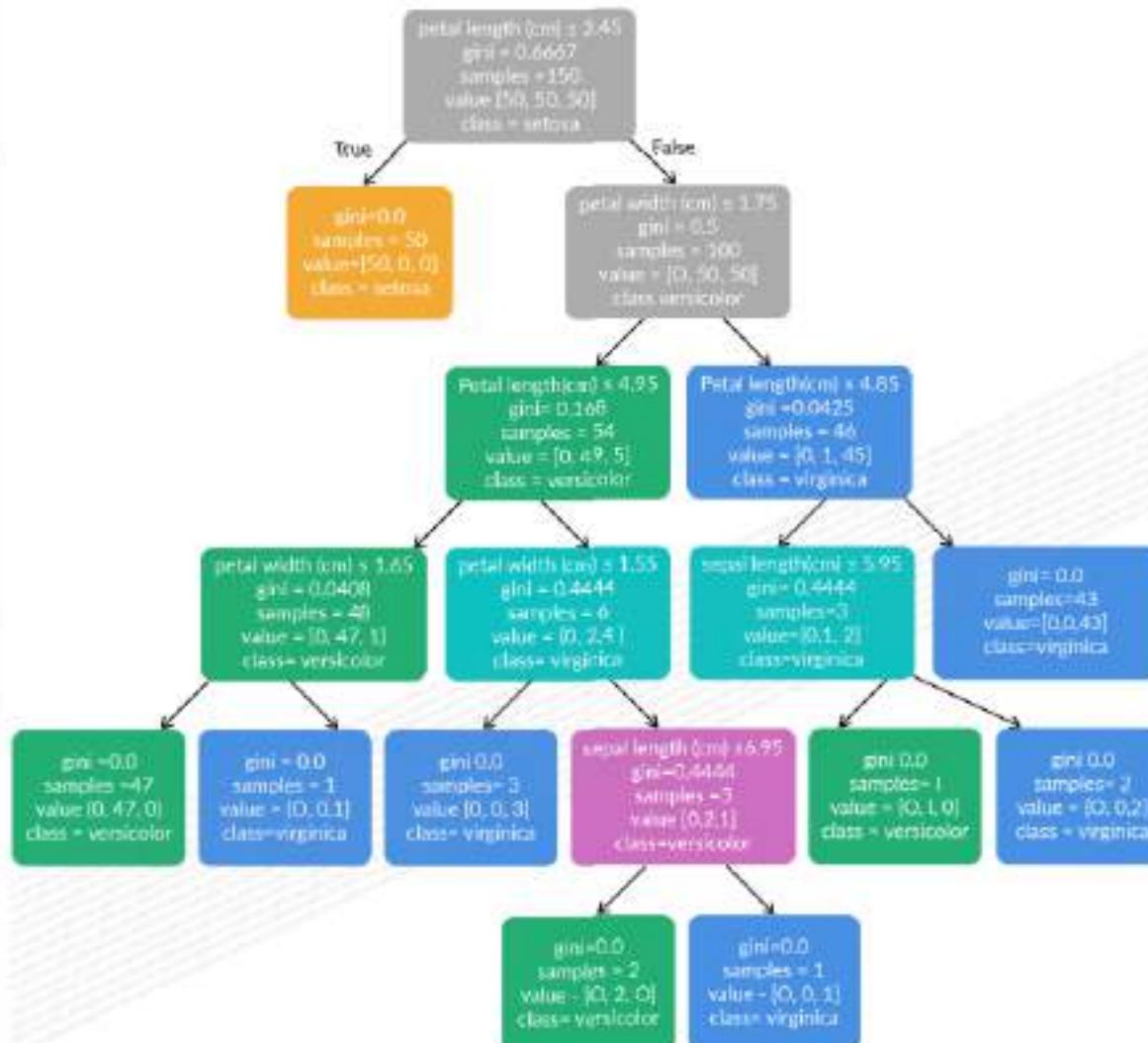
43 44 45

46 47 48

49 50 51

52 53 54

Note: In this decision tree you can't see all of the subtrees. That is because there are more than 100 nodes in this tree. This tree has approximately 100 nodes.



Answer Summary

Incorrect answers

- The given tree is overfitted tree. The tree looks above this tree has not good performance and accuracy on the test data.

- The tree is underfitted tree

- This tree gives lowest error than most of trees. It's a strong classifier. It's a good classifier.

- If the petal length is less than 2.45, then it is a good idea that the flower is not versicolor or virginica.

179 min left

9. C3 Advanced ML

[Previous](#)[Next](#)

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

Decision trees follow a top-down search as well as a greedy approach. What is the meaning of the term 'greedy' here?

Answer Options

Select any one option

[Clear Ans](#) It means that once a tree is created, we cannot add or delete a node. It means that a tree does not take into account what will happen in the next two or three steps. It means once a column is used for creating a split at a particular node, it can be used again for another split. It means that the tree cannot be pruned later.

49 50 51

52 53 54

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30



31 32 33

34 35 36

37 38 39

40 41 42

43 44 45

46 47 48

49 50 51

52 53 54

Refer to the table given below and find the OOB error for a random forest model that consists of four trees, A, B, C and D, and 10 observations. Use the table given below and fill the column 'Predicted Label', and then use it to calculate the OOB error.

Class Labels:

"O": Positive Class

"-": Negative Class

Observations	Tree Predictions				Actual Label	Predicted Label
	A	B	C	D		
2	O	-	O	O	O	
9	-	-	-	-	-	
10	O	O	O	-	O	
1	O	O	O	-	-	
7	-	-	O	-	-	
3	O	-	O	O	-	
6	-	-	-	-	-	
8	O	O	O	O	O	
4	-	-	-	O	O	
5	O	O	O	-	-	

Answer Options

Select any one option

Clear Ans

 0.2 0.4

179 min left

7. C3 Advanced ML

[Previous](#)[Next](#)

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30



31 32 33

34 35 36

37 38 39

40 41 42

43 44 45

46 47 48

49 50 51

52 53 54

Which of the following describes an advantage of the random forest (RF) algorithm over the decision tree (DT) algorithm?

Answer Options

Select any one option

[Clear All](#)

- An RF model has better interpretability as compared to a DT
- An optimally fit RF model has less variance than an optimally fit DT
- An optimally fit RF model has more variance than an optimally fit DT
- Building an RF model is computationally less expensive than building a DT

179 min left

6. C2 Clustering

[Previous](#)[Next](#)

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30



31 32 33

34 35 36

37 38 39

40 41 42

43 44 45

46 47 48

49 50 51

52 53 54

For a K-means clustering process, the Hopkins statistic for the dataset came out to be 0.8. Hence the dataset is

Answer Options

Select any one option

[Clear Ans](#) Suitable for clustering Not suitable for clustering Can't say from the given information None of the above

5. C2 Business Problem Solving

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

Any business problem solving will have following steps.(M)

1. To Identify the right data sources, that will be useful in formulating the final solution
2. Develop hypothesis and assess the overall impact of the hypothesized solution
3. Asking the right question for business and problem understanding
4. Define the solution approach: What will be the POC model? What will be the metrics for the model evaluation? etc.
5. Converting the business problem to a data science problem
6. Start your model building process with the simple POC model. And then increase the complexity of the POC model and optimize the parameters to get the best result.
7. Performing EDA on the datasets
8. Model Evaluation.

What will be the correct flow for solving the above/any business problem?

Answer Options

Select any one option

Clear Ans

 3>1>5>2>4>7>6>8 3>2>1>5>4>7>6>8 4>3>1>2>5>7>6>8 3>2>1>5>4>7>8>6

46 47 48

49 50 51

52 53 54

178 min left

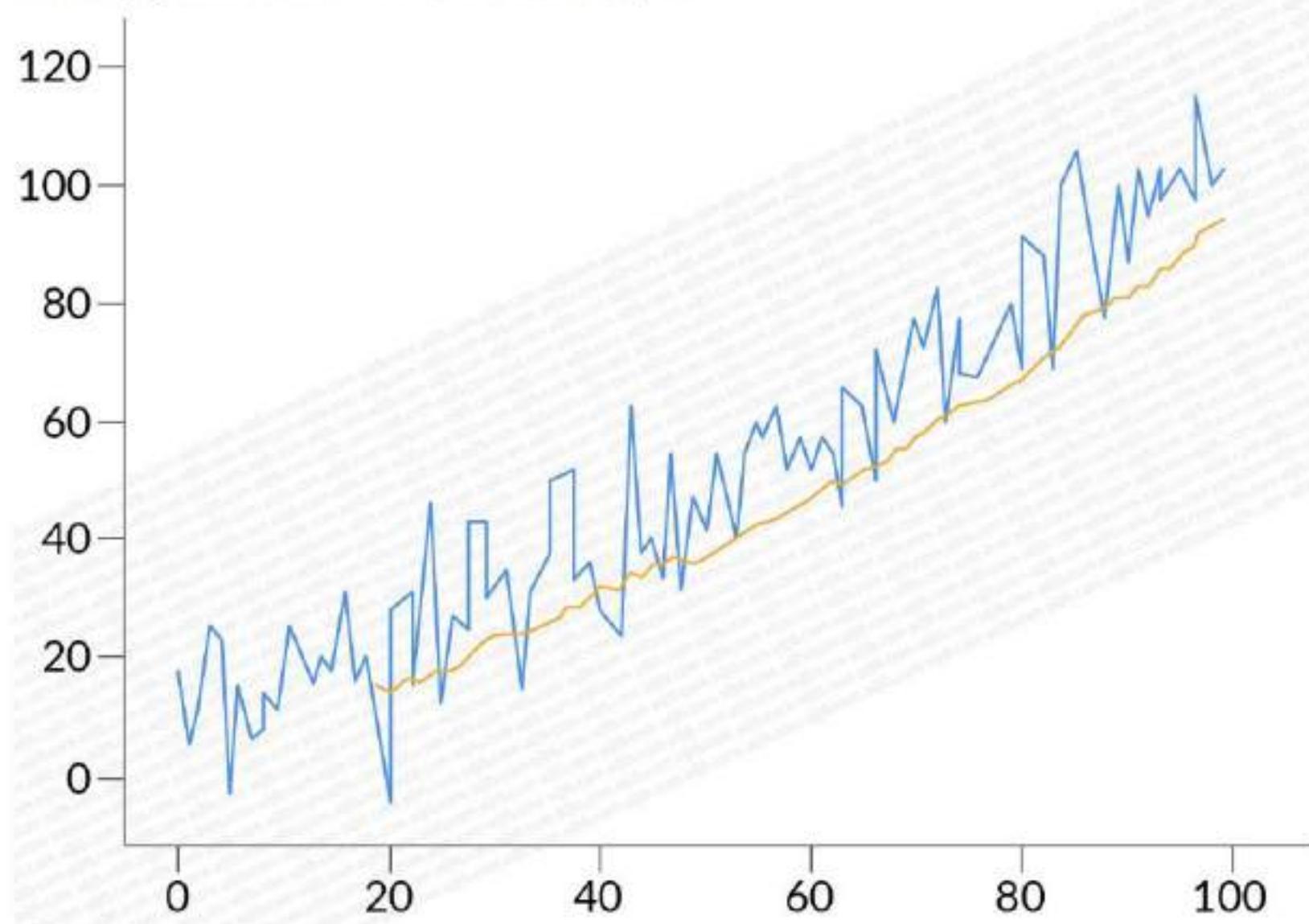
4. C3 Time Series Forecasting

< Previous

MCQs

1	2	3
4	5	6
7	8	9
10	11	12
13	14	15
16	17	18
19	20	21
22	23	24
25	26	27
28	29	30
31	32	33
34	35	36
37	38	39
40	41	42
43	44	45
46	47	48
49	50	51
52	53	54
55	56	57
58	59	60
61	62	63
64	65	66
67	68	69
70		

Refer to the image given below and determine whether the given time series is stationary or not.



Answer Options

Select any one option

C ✓

Coding

 Yes, it is stationary

Clear

3. C3 Time Series Forecasting

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

31 32 33

34 35 36

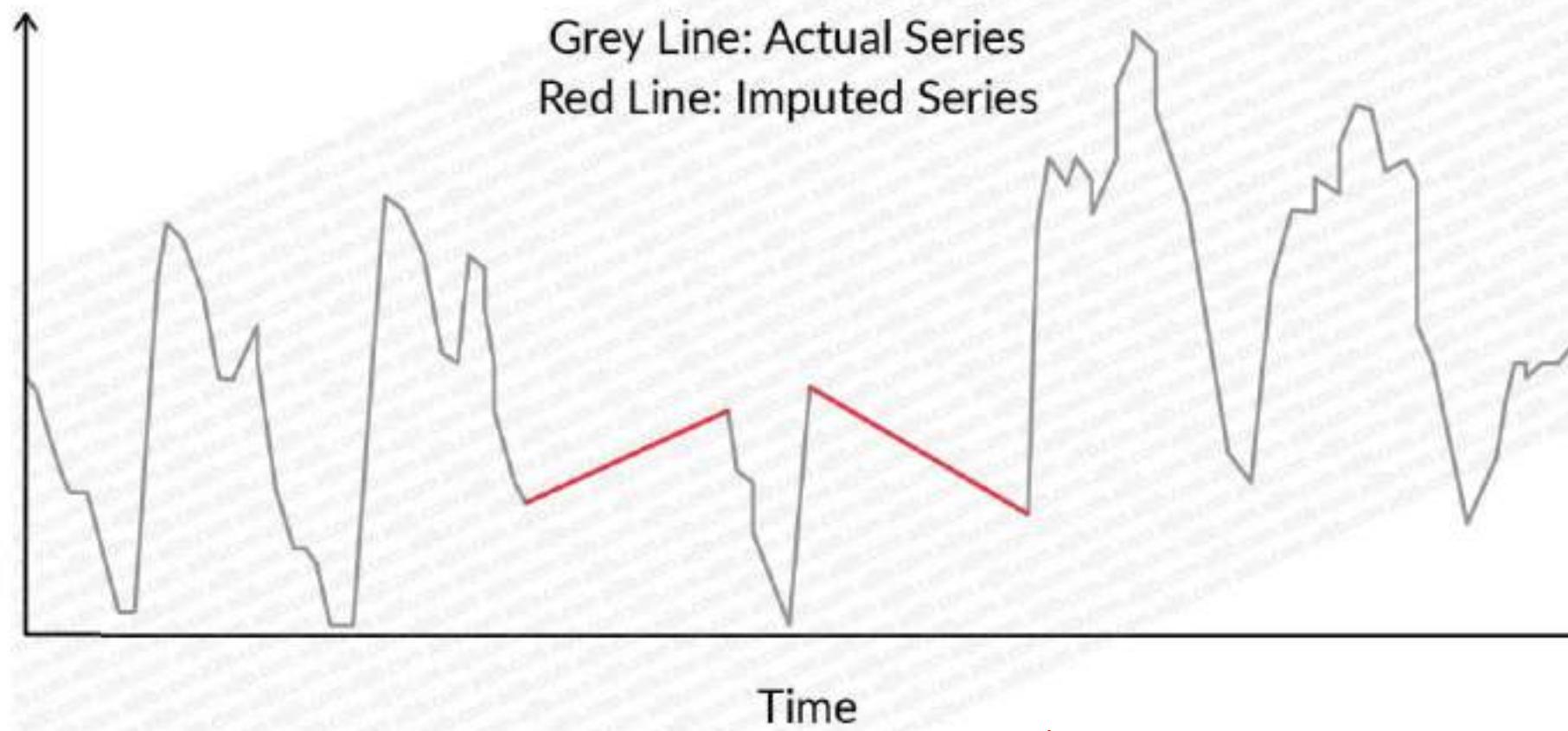
37 38 39

40 41 42

43 44 45

46 47 48

Refer to the image given below and choose the imputation method that is used to impute the missing time series.



Answer Options

Select any one option

Clear A

 Last observation carried forward Linear interpolation

180 min left

2. C3 Time Series Forecasting

[Previous](#)[Next](#)

MCQs

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8
- 9
- 10
- 11
- 12
- 13
- 14
- 15
- 16
- 17
- 18
- 19
- 20
- 21
- 22
- 23
- 24
- 25
- 26
- 27
- 28
- 29
- 30
- 31
- 32
- 33
- 34
- 35
- 36
- 37
- 38
- 39
- 40
- 41
- 42
- 43
- 44
- 45

Refer to the table given below and calculate the MAPE value

Year	Actual	Forecast
1990	12	14
1991	15	13
1992	13	13
1994	16	14

Answer Options

Select any one option

[Clear Ans](#) 15% 11%  8% 2%

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

Refer to the table below. The table consists of the Gini Impurity values before and after the split is made for various trees in a Random Forest model for the variable "Age". Calculate the feature importance of the variable "Age".

Random Forest			
Tree Number	Gini Impurity Before Split	Gini Impurity After Split	
1	0.48	0.32	
2	0.42	0.25	
3	0.44	0.13	
4	0.45	0.41	

Answer Options

25 **26** **27** **28** Select any one option

ClearAns

28 29 30 0.41

31 32 33

34 35 36

37 38 39

40 41 42

43 44 45

10 068

✓ 0.17

100 min left

4. Joins on MySQL

< Previous

Next >

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

31 32 33

34 35 36

37 38 39

40 41 42



43 44 45

category

Name	Type	Description
Id	int	
Name	varchar(20)	

customer_info

Name	Type	Description
Id	int	
City	varchar(25)	
First_Name	varchar(15)	
Last_Name	varchar(15)	
Segment	varchar(15)	
State	varchar(2)	
Street	varchar(40)	
Zipcode	varchar(5)	

product_info

Name	Type	Description
Product_Id	int	
Product_Name	varchar(50)	
Category_Id	int	
Department_Id	int	
Product_Price	decimal(12)	

ordered_items

Name	Type	Description
Order_Item_Id	int	
Order_Id	int	

SQLITE

1 -- Enter your query here

Coding

189 min left

4. Joins on MySQL

< Previous

Next >

SQLITE

1 --- Enter your query here

MCQs

1	2	3
4	5	6
7	8	9
10	11	12
13	14	15
16	17	18
19	20	21
22	23	24
25	26	27
28	29	30
31	32	33
34	35	36
37	38	39
40	41	42
43	44	45
46	47	48
49	50	51
52	53	54
55	56	57
58	59	60
61	62	63
64	65	66
67	68	69
70		

Order_

Table structure

department

Name	Type	Description
Id	int	
Name	varchar(20)	

category

Name	Type	Description
Id	int	
Name	varchar(20)	

customer_info

Name	Type	Description
Id	int	
City	varchar(25)	
First_Name	varchar(15)	
Last_Name	varchar(15)	
Segment	varchar(15)	
State	varchar(2)	
Street	varchar(40)	
Zipcode	varchar(5)	

product_info

Name	Type	Description
Product_Id	int	
Product_Name	varchar(50)	
Category_Id	int	

Coding

100 min left

4. Joins on MySQL

[◀ Previous](#)[Next ▶](#)

MCQs

Schema

```

1   2   3
4   5   6
7   8   9
10  11  12
13  14  15
16  17  18
19  20  21
22  23  24
25  26  27
28  29  30
31  32  33
34  35  36
37  38  39
40  41  42 <<
43  44  45
46  47  48
49  50  51
52  53  54
55  56  57
58  59  60
61  62  63
64  65  66
67  68  69
70
    
```

```

CREATE TABLE department (
    `Id` int NOT NULL,
    `Name` varchar(20) DEFAULT NULL,
    PRIMARY KEY (`Id`)
)

CREATE TABLE category (
    `Id` int NOT NULL,
    `Name` varchar(20) DEFAULT NULL,
    PRIMARY KEY (`Id`)
)

CREATE TABLE customer_info (
    `Id` int NOT NULL,
    `City` varchar(25) DEFAULT NULL,
    `First_Name` varchar(15) DEFAULT NULL,
    `Last_Name` varchar(15) DEFAULT NULL,
    `Segment` varchar(15) DEFAULT NULL,
    `State` varchar(2) DEFAULT NULL,
    `Street` varchar(40) DEFAULT NULL,
    `Zipcode` varchar(5) DEFAULT NULL,
    PRIMARY KEY (`Id`)
)

CREATE TABLE product_info (
    `Product_Id` int NOT NULL,
    `Product_Name` varchar(50) DEFAULT NULL,
    `Category_Id` int DEFAULT NULL,
    `Department_Id` int DEFAULT NULL,
    `Product_Price` decimal(12, 2) DEFAULT NULL,
    PRIMARY KEY (`Product_Id`)
)

CREATE TABLE ordered_items (
    `Order_Item_Id` int NOT NULL,
    `Order_
```

Table structure

department

Name	Type	Description
Id	Int	
Name	varchar(20)	

Coding

SQLITE

1 -- Enter your query here

100 min left

4. Joins on MySQL

< Previous

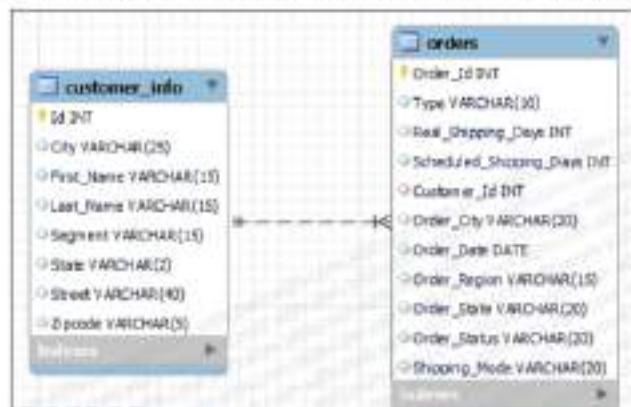
Next >

MDQs

1 2 3
4 5 6
7 8 9
10 11 12
13 14 15
16 17 18
19 20 21
22 23 24
25 26 27
28 29 30
31 32 33
34 35 36
37 38 39
40 41 42
43 44 45
46 47 48
49 50 51
52 53 54
55 56 57
58 59 60
61 62 63
64 65 66
67 68 69
70

INPUT TABLE

You are given two tables - customer_info and orders whose schema is given below:



QUERY:

Display all the details (all columns) from the **orders** table
where

- * State is AZ and
- * Street contains the word 'Silver'

Order them by Order_Id in ascending order.

Note – The Id column in customer_info is common with the Customer_Id column in the orders table.

OUTPUT COLUMNS:

All column of orders table

Note that the coding console automatically converts the casing of the columns to upper case

PLEASE NOTE:

- * Use the original column names only. Any other aliases or column names would lead to an error.
- * Keep the sequence of the columns same as in the original table.

Schema

Table structure

department

Name	Type	Description
id	int	
Name	varchar(20)	

Coding

SQLITE

1 -- Enter your query here

100 min left

3. kth Largest

[Previous](#)[Next](#)

MCQs

1 2 3
4 5 6

7 8 9
10 11 12

13 14 15
16 17 18

19 20 21
22 23 24

25 26 27
28 29 30

31 32 33
34 35 36

37 38 39
40 41 42

43 44 45
46 47 48

49 50 51
52 53 54

55 56 57
58 59 60

61 62 63
64 65 66

67 68 69
70

Input Format:

Line 1 contains a list of integers
Line 2 contains a positive integer $k > 0$

Output Format:

An integer, k th largest integer

Examples:

Sample Input 1:

[2, 3, 1, 5, 6, 2, 1]

4

Sample Output 1:

2

Sample Input 2:

[2, 3, 1, 5, 6, 2, 1]

6

Sample Output 2:

-1

Function description

Complete the `kthLargest` function in the editor below. It has the following parameter(s):

Name	Type	Description
<code>k</code>	INTEGER	
<code>arr</code>	INTEGER ARRAY	

Return

The function must return an INTEGER denoting the k th largest integer in the array

Constraints

- $1 \leq m \leq 10^5$
- $1 \leq k \leq 10^5$
- $1 \leq arr[i] \leq 10^5$

Input format for debugging

- The first line contains an integer array of m , denoting the number of elements in `arr`.
- The next line contains an integer, k , denoting the k th largest number that the function should return.
- Each line i of the m subsequent lines (where $0 \leq i < m$) contains an integer describing `arr[i]`.

Sample Testcases

Python 3

```

1 import sys
2
3
4 def kthLargest(arr, k):
5     # Write your code here
6
7
8
9 def main():
10    import ast
11    arr = []
12    arr = ast.literal_eval(input())
13    k = int(input())
14    result = kthLargest(arr, k)
15    print(result)
16
17
18 if __name__ == '__main__':
19    main()

```

Coding

Input

Output

Output Description

Console

Custom Test Case

100 min left

3. kth Largest

< Previous

Next >

MOUs

1 2 3
4 5 6
7 8 910 11 12
13 14 15
16 17 1819 20 21
22 23 24
25 26 2728 29 30
31 32 33
34 35 3637 38 39
40 41 42
43 44 45

Problem Statement

Given a list of integers and another integer k , find the k th largest integer in the list. If there are less than k distinct elements in the list, then you need to output -1. (Refer to the sample inputs and outputs for details.)

Note: $k=1$ means you have to find the largest element

Input Format:

Line 1 contains a list of integers
Line 2 contains a positive integer $k > 0$

Output Format:

An Integer, k th largest integer

Examples:

Sample Input 1:

[2, 3, 1, 5, 0, 2, 1]

4

Sample Output 1:

2

Sample Input 2:

[2, 3, 1, 5, 0, 2, 1]

6

Sample Output 2:

-1

Function description

Complete the `kthLargest` function in the editor below. It has the following parameter(s):

Name	Type	Description
<code>k</code>	INTEGER	
<code>arr</code>	INTEGER ARRAY	

Return The function must return an INTEGER denoting the k th largest integer in the array.

Constraints

Input format for debugging

Sample Testcases

Python 3

```

1 import sys
2
3
4 def kthLargest(arr, k):
5     # Write your code here
6
7
8
9 def main():
10    import ast
11    arr = []
12    arr = ast.literal_eval(input())
13    k = int(input())
14    result = kthLargest(arr, k)
15    print(result)
16
17
18 if __name__ == '__main__':
19    main()

```

170 min left

2. GroupBy

[◀ Previous](#)[Next ▶](#)

SQLITE

1 -- Enter your query here

MGR_ID

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

31 32 33

34 35 36

37 38 39

40 41 42

Name	Type	Description
Id	int	
Name	varchar(20)	

customer_info

Name	Type	Description
Id	int	
City	varchar(25)	
First_Name	varchar(15)	
Last_Name	varchar(15)	
Segment	varchar(15)	
State	varchar(2)	
Street	varchar(40)	
Zipcode	varchar(5)	

product_info

Name	Type	Description
Product_Id	int	
Product_Name	varchar(50)	
Category_Id	int	
Department_Id	int	
Product_Price	decimal(12)	

ordered_items

Name	Type	Description
OrderItem_Id	int	
Order_Id	int	
Item_Id	int	

Coding

Console Custom Test Case

120 min left

2. GroupBy

[◀ Previous](#)[Next ▶](#)

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

31 32 33

34 35 36

37 38 39

40 41 42

43 44 45

46 47 48

49 50 51

52 53 54

55 56 57

58 59 60

61 62 63

64 65 66

67 68 69

70

Coding

* Order_

Table structure

department

Name	Type	Description
id	int	
Name	varchar(20)	

category

Name	Type	Description
id	int	
Name	varchar(20)	

customer_info

Name	Type	Description
id	int	
City	varchar(25)	
First_Name	varchar(15)	
Last_Name	varchar(15)	
Segment	varchar(15)	
State	varchar(2)	
Street	varchar(40)	
Zipcode	varchar(5)	

product_info

Name	Type	Description
Product_Id	int	
Product_Name	varchar(50)	
Category_Id	int	

SQLITE

1 -- Enter your query here

129 min left

2. GroupBy

< Previous

Next >

SQLITE

1 --- Enter your query here

MCQs

1	2	3
4	5	6
7	8	9
10	11	12
13	14	15
16	17	18
19	20	21
22	23	24
25	26	27
28	29	30
31	32	33
34	35	36
37	38	39
40	41	42
43	44	45
46	47	48
49	50	51
52	53	54
55	56	57
58	59	60
61	62	63
64	65	66
67	68	69
70		

PLEASE NOTE:

- Use the alias of 'oc' for the count of orders. Any other alias or column name would lead to an error.
- Please keep the sequence same as mentioned in the output columns. For example in the above scenario, if you display Type,oc instead of oc,Type you'll get an error.

Schema

```

CREATE TABLE department (
    `Id` int NOT NULL,
    `Name` varchar(20) DEFAULT NULL,
    PRIMARY KEY (`Id`)
)

CREATE TABLE category (
    `Id` int NOT NULL,
    `Name` varchar(20) DEFAULT NULL,
    PRIMARY KEY (`Id`)
)

CREATE TABLE customer_info (
    `Id` int NOT NULL,
    `City` varchar(25) DEFAULT NULL,
    `First_Name` varchar(15) DEFAULT NULL,
    `Last_Name` varchar(15) DEFAULT NULL,
    `Segment` varchar(15) DEFAULT NULL,
    `State` varchar(2) DEFAULT NULL,
    `Street` varchar(40) DEFAULT NULL,
    `Zipcode` varchar(5) DEFAULT NULL,
    PRIMARY KEY (`Id`)
)

CREATE TABLE product_info (
    `Product_Id` int NOT NULL,
    `Product_Name` varchar(50) DEFAULT NULL,
    `Category_Id` int DEFAULT NULL,
    `Department_Id` int DEFAULT NULL,
    `Product_Price` decimal(12, 2) DEFAULT NULL,
    PRIMARY KEY (`Product_Id`)
)

CREATE TABLE ordered_items (
    `Order_Item_Id` int NOT NULL,
    `Order_`
```

Table structure

department

Name

Description

Coding

Console

Custom Table Case

120 min left

2. GroupBy

© Prentice

Next →

1472

1	2	3
4	5	6
7	8	9
10	11	12
13	14	15
16	17	18
19	20	21
22	23	24

28 29 30

31 32 33
34 35 36
37 38 39
40 41 42

42 44 45

卷首 47 48

40 50 51

53 54

55 56 57

549

• 100 •

INPUT TABLE

You're given a `orders` table and the columns in the `orders` table are shown below:

orders	
Order_Id	INT
Type	VARCHAR(10)
Real_Shipping_Days	INT
Scheduled_Shipping_Days	INT
Customer_Id	INT
Order_City	VARCHAR(20)
Order_Date	DATE
Order_Region	VARCHAR(10)
Order_Status	VARCHAR(10)
Shipping_Mode	VARCHAR(20)

CHERRY

- Calculate count of all the orders.
 - Where Order_State is Maharashtra
 - Note - Use the alias of `oc` for count of orders.
 - Group the results by Type
 - Order them by `oc` in ascending order.

DISTRIBUTION COLUMNS

CH. THREE

Here's an image showing how a sample output would look like:

OC	TYPE
45	CASH
69	PAYMENT
108	TRANSFER
151	DEBIT

Note that the coding console automatically converts the casing of the columns to upper case.

RELEASE NOTE

- Use the alias of 'oc' for the count of orders. Any other alias or column name would lead to an error.
 - Please keep the sequence same as mentioned in the output columns. For example in the above scenario, if you display `Type,oc` instead of `oc,Type` you'll get an error.

© Cambridge

150/178

1 — ESTCE VRAI QU'IL FAUT

170 min left

1. GroupBy and OrderBy

< Previous

Next >

SQLITE

1 --- Enter your query here

MCQs

1 2 3
4 5 6
7 8 9
10 11 12
13 14 15
16 17 18
18 19 20
22 23 24
26 26 27
28 29 30
31 33 33
34 35 36
37 38 39
41 41 42
43 44 45
46 47 48
50 50 51
53 53 54
56 56 57
58 59 60
61 62 63
64 65 66
67 68 69
70

Product_Price	decimal(12)	
ordered_items		
Name	Type	Description
Order_Item_Id	int	
Order_Id	int	
Item_Id	int	
Quantity	int	
Sales	decimal(12)	

Sample testcase 1

Input

department

id	Name
----	------

category

id	Name
----	------

customer_info

id	City	First_Name	Last_Name	Segment	State	Street	Zipcode
----	------	------------	-----------	---------	-------	--------	---------

product_info

Product_id	Product_Name	Category_Id	Department_Id	Product_Price
------------	--------------	-------------	---------------	---------------

OrderItem

Order_Item_Id	Order_Id	Item_Id	Quantity	Sales
---------------	----------	---------	----------	-------

Output

1	Jamnagar
1	Rajkot
1	Vadodara

Coding

1 2 3
4 5 6
7 8 9
10 11 12
13 14 15
16 17 18
18 19 20
22 23 24
26 26 27
28 29 30
31 33 33
34 35 36
37 38 39
41 41 42
43 44 45
46 47 48
50 50 51
53 53 54
56 56 57
58 59 60
61 62 63
64 65 66
67 68 69
70

Console

Custom Test Cases

129 min left

1. GroupBy and OrderBy

< Previous Next >

SQLITE

1 --- Enter your query here

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

26 28 27

28 29 30

31 32 33

<<

Table structure

department

Name	Type	Description
Id	int	
Name	varchar(20)	

category

Name	Type	Description
Id	int	
Name	varchar(20)	

customer_info

Name	Type	Description
Id	int	
City	varchar(25)	
First_Name	varchar(15)	
Last_Name	varchar(15)	
Segment	varchar(15)	
State	varchar(2)	
Street	varchar(40)	
Zipcode	varchar(5)	

product_info

Name	Type	Description
Product_Id	int	
Product_Name	varchar(50)	
Category_Id	int	

Coding

1 2 3

4 5 6

7 8 9

Console Custom Test Case

171 min left

1. GroupBy and OrderBy

< Previous Next >

SQLITE

1 --- Enter your query here

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

31 32 33

34 35 36

37 38 39

40 41 42

43 44 45

46 47 48

49 50 51

52 53 54

55 56 57

58 59 60

61 62 63

64 65 66

67 68 69

Schema

```

CREATE TABLE department (
    `Id` int NOT NULL,
    `Name` varchar(20) DEFAULT NULL,
    PRIMARY KEY (`Id`)
)

CREATE TABLE category (
    `Id` int NOT NULL,
    `Name` varchar(20) DEFAULT NULL,
    PRIMARY KEY (`Id`)
)

CREATE TABLE customer_info (
    `Id` int NOT NULL,
    `City` varchar(25) DEFAULT NULL,
    `First_Name` varchar(15) DEFAULT NULL,
    `Last_Name` varchar(15) DEFAULT NULL,
    `Segment` varchar(15) DEFAULT NULL,
    `State` varchar(2) DEFAULT NULL,
    `Street` varchar(40) DEFAULT NULL,
    `Zipcode` varchar(5) DEFAULT NULL,
    PRIMARY KEY (`Id`)
)

CREATE TABLE product_info (
    `Product_Id` int NOT NULL,
    `Product_Name` varchar(50) DEFAULT NULL,
    `Category_Id` int DEFAULT NULL,
    `Department_Id` int DEFAULT NULL,
    `Product_Price` decimal(12, 2) DEFAULT NULL,
    PRIMARY KEY (`Product_Id`)
)

CREATE TABLE ordered_items (
    `Order_Item_Id` int NOT NULL,
    `Order_`
```

Coding

Table structure

171 min left

1. GroupBy and OrderBy

Previous Next

SQLITE

1 -- Enter your query here

MCQs

1	2	3
4	5	6
7	8	9
10	11	12
13	14	15
16	17	18
19	20	21
22	23	24
25	26	27
28	29	30

INPUT TABLE

You're given a **orders** table and the columns in the orders table are shown below:

orders		
Order_Id	INT	
Type	VARCHAR(10)	
Real_Shipping_Days	INT	
Scheduled_Shipping_Days	INT	
Customer_Id	INT	
Order_City	VARCHAR(20)	
Order_Date	DATE	
Order_Region	VARCHAR(15)	
Order_Status	VARCHAR(20)	
Shipping_Mode	VARCHAR(20)	

QUERY

- Calculate count of all the orders
 - where the Order_Status is **Gujarat**
- where the Order_Status is **PENDING**.
 - Note - Use the alias of **oc** for count of orders.
- Group the results by Order_City
- Order them by **oc** & Order_City in **ascending order**.

OUTPUT COLUMNS

oc, Order_City

Here's an image showing how a sample output would look like. :

OC	ORDER_CITY
1	Jamnagar
1	Rajkot
1	Vadodara
2	Jodhpur
3	Bhavnagar
4	Surat

Note that the coding console automatically converts the casing of the columns to upper case

Coding

PLEASE NOTE

Console

Output Test Case

172 min left

70. C3 Time Series Forecasting

< Previous

Next >

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

31 32 33

34 35 36

37 38 39

40 41 42

43 44 45

46 47 48

49 50 51

52 53 54

55 56 57

58 59 60

61 62 63

64 65 66

67 68 69

ACF values of a time series of length 100 computed over 20 lag values are -0.1, -0.15, 0.06, 0.20, -0.19, 0.16, -0.08, 0.21, -0.11, -0.09, 0.18, 0.16, 0.18, -0.22, 0.10, -0.08, 0.07, -0.09, 0.15, -0.16. How many of the values are outside the confidence limits at a 95% p-value threshold?
(Hint: Interval is given by $-Z^*/\sqrt{\text{length}}$, $Z^*/\sqrt{\text{length}}$, Z^* at 95% = 1.96)

?

Answer Options

Select any one option

Clear

 5 0 3 None of the above

172 min left

69. C3 Multiple Correct Answer

< Previous

MCQs

- 1 2 3
- 4 5 6
- 7 8 9
- 10 11 12
- 13 14 15
- 16 17 18
- 19 20 21

- 22 23 24

How is diversity achieved in the case of a random forest? (Note: Multiple options might be correct)

- 25 26 27

Select one or more options.

Clear

We perform tree pruning for each tree that is built on the bootstrapped sample.

Not all the features are considered at each node of the tree that is built on the bootstrapped sample; rather, a subset of random features is considered.

Bootstrapped sampling is performed over the data to create multiple samples.

A random subset of features is considered for building each tree.

- 48 49 50

- 52 53 54

- 56 57 58

- 60 61 62

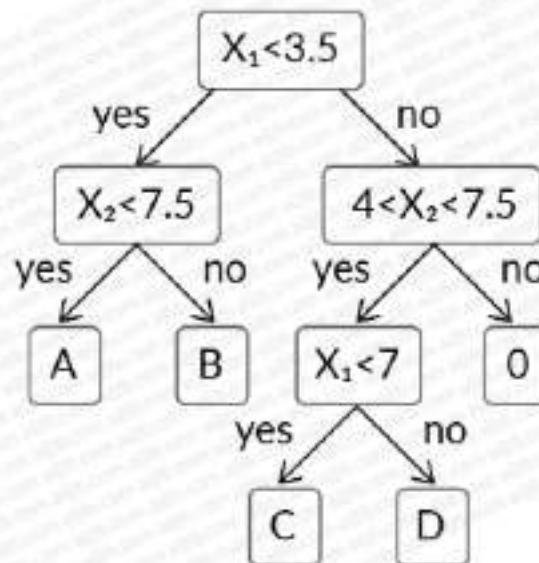
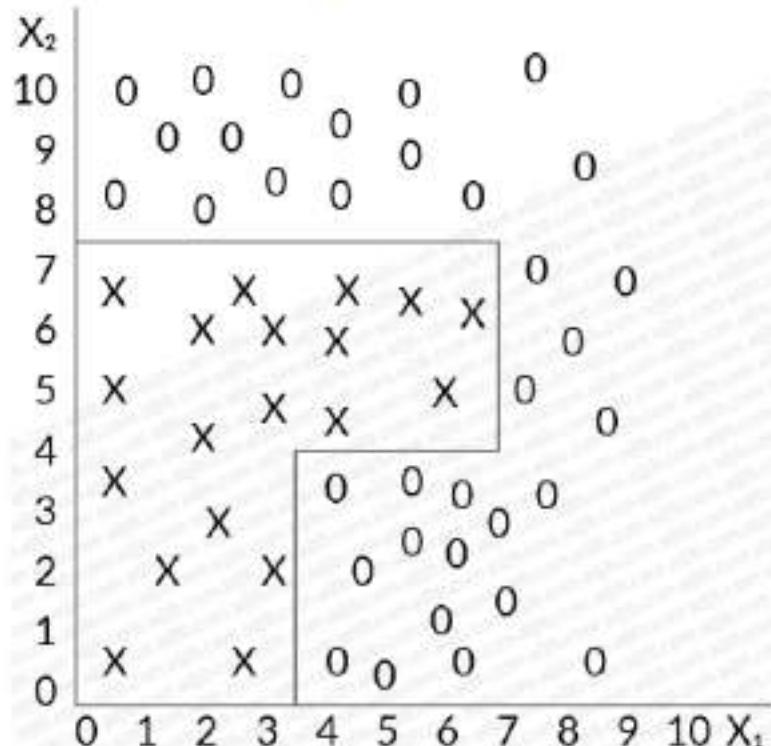
- 64 65 66

- 67 68 69

MCQs

- 1 2 3
4 5 6
7 8 9
10 11 12
13 14 15
16 17 18
18 19 20
20 21
22 23 24
23 26 27
28 29 30
31 32 33
34 35 36
37 38 39
40 41 42
42 44 45
46 47 48
49 50 51
52 53 54
55 56 57
58 59 60
61 62 63
64 65 66
67 68 69
70

Refer to the image with the decision tree and the boundary diagram. Find what will be the outcome for the leaf node: A, B, C, D.
Note: All outcome can be an "o" or an "x".



Answer Options

Select any one option

Clear

 A: x, B: x, C: o, D: o A: o, B: o, C: x, D: x A: x, B: o, C: x, D: o

172 min left

67. C2 Logistic Regression

< Previous

MCQs

- 1 2 3
- 4 5 6
- 7 8 9
- 10 11 12
- 13 14 15
- 16 17 18
- 19 20 21

22. 23. 24. Answer Options

Select any one option

Clear

- It helps in capturing the seasonal fluctuations that might be present in the data
- It helps to find the optimal cutoff point more easily
- It helps in finding the different predictive patterns for the different set of data points that might be present in the data
- It helps capture the trends easily when there is a class imbalance

- 65 66 67
- 68 69 70
- 71 72 73
- 74 75 76
- 77 78 79

Coding

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

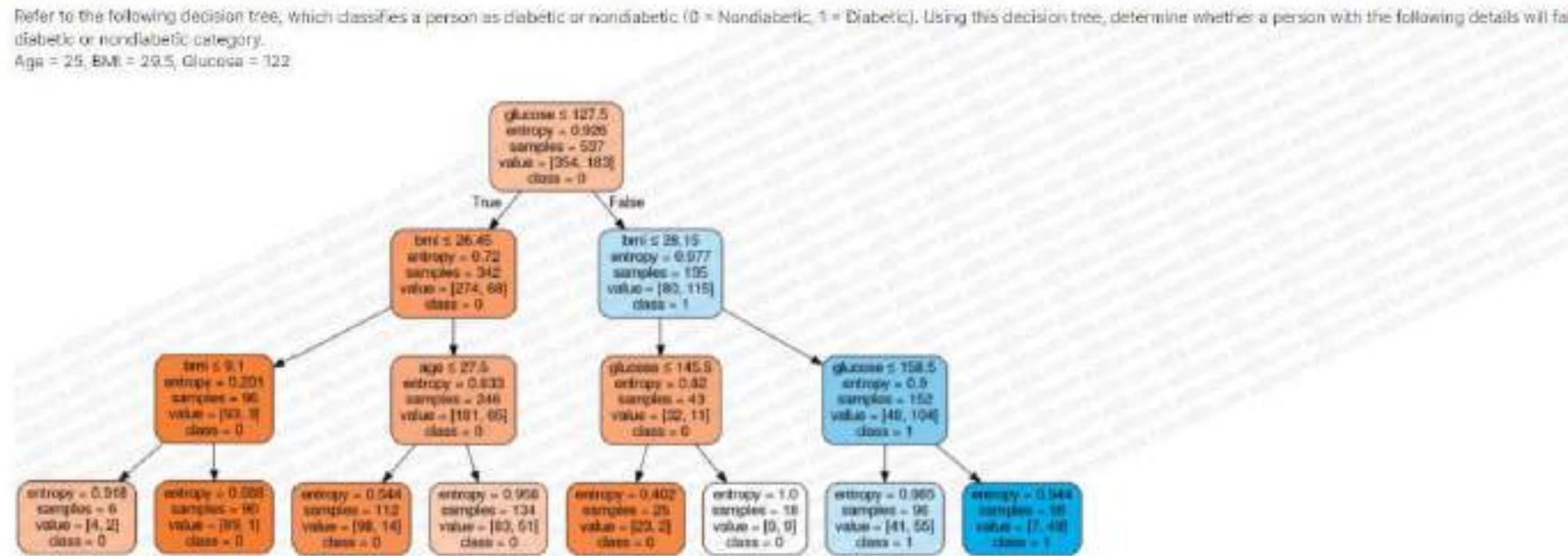
28 29 30

31 32 33

34 35 36

37 38 39

40 41 42



Answer Options:

Select any one option

Clear

 Diabetic Non-diabetic Insufficient information

172 min left

65. C2 Multiple correct answer

< Previous

MCQs

- 1 2 3
- 4 5 6
- 7 8 9
- 10 11 12

The ROC curve shows the tradeoff between the True Positive Rate (TPR) and the False Positive Rate (FPR). The following function is written in Python using the metrics package from the scikit-learn library. The ROC curve function.

```
def draw_roc(actual, probs):  
    fpr, tpr, thresholds=metrics.roc_curve(actual,probs,drop_intermediate=False)  
    auc_score = metrics.roc_auc_score(actual,probs)  
    return None
```

Which of the following statements are true? (More than one option may be correct.)

- 13 14 15
- 16 17 18
- 19 20 21

Answer Options

- 25 26 27

Select one or more options.

Clear

The arguments passed in the above function are actual values of the target variable and the predicted values (i.e., 0 or 1)

The arguments passed in the above function are actual values of the target variable and the respective predicted probabilities

The area under the ROC can take any value between 0 and 1

Larger the area under the curve, the better will be the model

- 52 53 54
- 55 56 57
- 58 59 60

- 61 62 63

- 64 65 66

- 67 68 69

122 min left

64. C2 Multiple correct answer

[Previous](#)

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

26 28 27

28 29 30

31 32 33

34 35 36

37 38 39

41 42

43 44 45

46 47 48

50 51

53 54

56 58 57

58 59

61 62 63

64 65 66

67 68 69

70

Coding

Which of the following metrics can be used for finding the appropriate number of clusters in K-means clustering? (More than one option may be correct)

Answer Options

Select one or more options

[Clear](#) Silhouette Score Elbow Curve Hopkins Statistic Dendogram

172 min left

63. C2 Multiple correct answer

[Previous](#)

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

26 28 27

28 29 30

31 32 33

34 35 36

37 38 39

41 42

43 44 45

46 47 48

50 51

53 54

56 58 57

58 59

61 62

64 65 66

67 68 69

70

Which of the following statements are correct in the context of logistic regression? (More than one option may be correct.)

Answer Options

Select one or more options

[Clear](#) The dummies for continuous variables make the model more unstable Weight of evidence (WoE) helps in treating missing values for both continuous and categorical variables WoE should follow a non-monotonic trend across bins. Data clumping can be a problem with transforming continuous variables to dummies. Information value or IV is an important indicator of predictive power.

122 min left

62. C2 Logistic Regression

< Previous

Next >

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

31 32 33

34 35 36

37 38 39

40 41 42

43 44 45

46 47 48

49 50 51

52 53 54

55 56 57

58 59 60

61 62 63

64 65 66

67 68 69

70

Take a look at the following three problem statements.

problem statement 1: Let's say that you are building a telecom churn prediction model with the business objective that your company wants to implement an aggressive customer retention campaign to retain 'high churn-risk' customers. This is because a competitor has launched extremely low-cost mobile plans, and you want to avoid churn as much as possible by incentivising the customers. Assume that budget is not a constraint.

problem statement 2: Let's say you are building a cancer detection model with the objective that both the patient who has cancer and the patient who has not cancer can be detected correctly. It can have serious implications if you predict either of the class wrong. i.e., If wrongly detected as 'not cancer', the patient will die of cancer, and if wrongly detected as 'cancer', the patient will die of chemotherapy.

Problem statement 3: You have to build an image classification model where 60% of images belong to one class and rest 40% images belong to another class. You have to predict the class of a new image.

Which is the correctly matched model evaluation metric for the above classification models?

Answer Options

Select any one option

Clear

 Problem statement 1: Specificity Problem statement 2: Sensitivity Problem statement 2: Specificity Problem statement 3: Accuracy

Coding

172 min left

61. C2 linear Regression

< Previous

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

31 32 33

34 35 36

37 38 39

40 41 42

43 44 45

46 47 48

49 50 51

52 53 54

55 56 57

58 59 60

61 62 63

64 65 66

67 68 69

70

Coding

RFE method is used for:

Answer Options

Select any one option

Clear

 Dummy variable creation Detecting multicollinearity Feature selection Univariate regression

174 min left

60. C2 linear Regression

< Previous

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

Answer Options

25 26 27

Select any one option

Clear

 Mean of residuals of old model > Mean of residuals of new model. Mean of residuals of old model < Mean of residuals of new model. Mean of residuals of old model = Mean of residuals of new model. Information provided is not enough to comment on the mean of residuals.

59 60 61

62 63

64 65 66

67 68 69

70

60

174 min left

59. C2 Multiple correct answer

< Previous

MCQs

- 1 2 3
- 4 5 6
- 7 8 9
- 10 11 12
- 13 14 15
- 16 17 18
- 19 20 21
- 22 23 24

Which of the following statements are true? (More than one option may be correct.)

Answer Options

Select one or more options

Clear

TSS (Total Sum of Squares) is defined as the sum of all squared differences between the observed dependent variable and its mean.

R-squared can take any value between 0 and 1.

Larger the R-squared value, the better the regression model fits the observations.

If RSS = 5.50 and TSS = 11, the value of VIF will be 1.33.

<<

- 55 56 57
- 58 59 60
- 61 62 63
- 64 65 66
- 67 68 69
- 70

Coding

MCQs

1 2 3
4 5 6
7 8 9
10 11 12
13 14 15
16 17 18
19 20 21

22 23 24
25 26 27
28 29 30
31 32 33
34 35 36
37 38 39
40 41 42
43 44 45
46 47 48
49 50 51
52 53 54
55 56 57
58 59 60
61 62 63
64 65 66
67 68 69
70

We wanted to tune hyperparameters for a Random Forest model using Grid Search technique with the CV of 5 folds. Refer to the list of hyperparameters that are required to be tuned.

```
{'n_estimators': [10, 25], 'max_features': [5,10],  
 'max_depth': [10, 50, None], 'bootstrap': [True, False]}  
 ]
```

Assume that each set of hyperparameters takes 1 minutes to run, find the total time required to complete the tuning process.

Wrong options Ans. 120

Answer Options

Select any one option

30 minutes

48 minutes

72 minutes

96 minutes

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

31 32 33

34 35 36

37 38 39

40 41 42

43 44 45

46 47 48

49 50 51

52 53 54

57

56 57 58

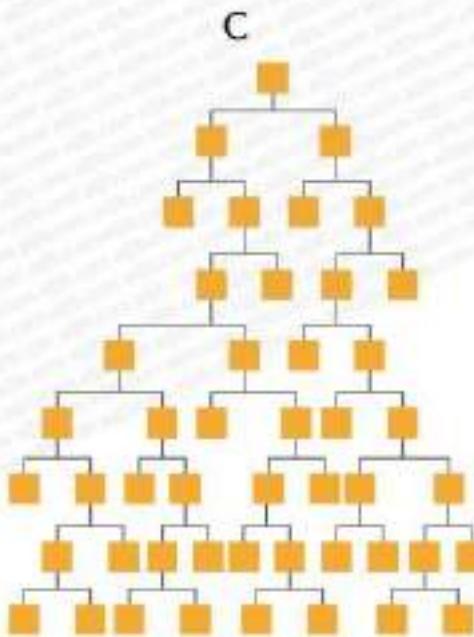
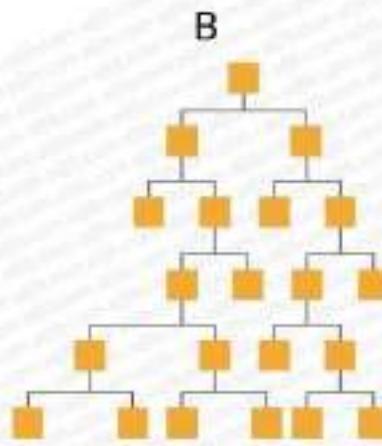
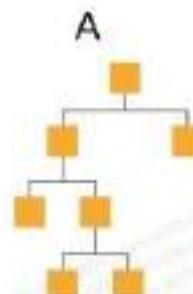
61 62 63

64 65 66

67 68 69

70

Which of the following models would have low bias but high variance?



Accuracy on training = 50%

Accuracy on test = 50%

Accuracy on training = 70%

Accuracy on test = 70%

Accuracy on training = 90%

Accuracy on test = 65%

C. ✓

Answer Options

Select any one option

A

Coding

MCQs:

- 1 2 3
- 4 5 6
- 7 8 9
- 10 11 12
- 13 14 15
- 16 17 18
- 19 20 21
- 22 23 24
- 25 26 27
- 28 29 30
- 31 32 33
- 34 35 36
- 37 38 39
- 40 41 42
- 43 44 45



Answer Options

Select any one option

Clear A

- The cross-validation scheme is computationally expensive.
- D. The scheme always generate low bias and low variance model
- It is better when you have a small set of training data.
- The model is fitted on $n - 1$ training samples under this cross-validation scheme.

174 min left

55. C3 Advanced ML

< Previous

Next

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30



31 32 33

34 35 36

37 38 39

40 41 42

43 44 45

46 47 48

49 50 51

52 53 54

Which of the following is the correct sampling technique that is used by a random forest model to overcome the problem of overfitting?

Answer Options

Select any one option

Clear Ans

 Random sampling Bootstrapping Oversampling Stratified sampling

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30



31 32 33

34 35 36

37 38 39

40 41 42

43 44 45

46 47 48

49 50 51

52 53 54

The coronavirus disease (COVID-19), was declared a pandemic by the World Health Organization (WHO) in February 2020. Currently, there are no vaccines or treatments that have been officially approved by WHO after clinical trials. India has not seen the peak of infection yet and the number of infections is touching a new height daily. The business unit of an Indian health and hygiene company approaches you to know "**Why the sales of masks is decreasing despite the number of corona infections increasing daily**".

Answer the below question:

Consider the following two statements:

Statement 1: Understanding the change in customer behaviour is an important factor to be considered for business understanding for the problem statement above.

Statement 2: One of the possible hypotheses for the above problem statement: There is a rise in the number of companies manufacturing normal/surgical masks due to which the sales of the client's company is decreasing.

Answer Options

Select any one option

Clear Ans

Statement 1 is correct and Statement 2 is wrong

Statement 2 is correct and Statement 1 is wrong

Both the statements are correct

None of the statements are correct

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

<<

31 32 33

34 35 36

37 38 39

40 41 42

43 44 45

46 47 48

49 50 51

52 53 54

What would happen if you choose a very high value of the hyperparameter, lambda?

$$\min_{a,b} \left[\sum_{i=1}^n (y_i - ax_i - b)^2 + \lambda(a^2 + b^2) \right]$$

Answer Options

Select any one option

Clear Ans

 The model would become simpler, yet it would show robust performance on test data. The model would become too simple. It would become an underfitted model. The model would become too complex. It would become an overfitted model. The model would show good performance on the train data, but it will show poor performance on the test data.

174 min left

52. C2 linear Regression

[Previous](#)[Next](#)

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

Answer Options

Select any one option

[Clear All](#) The r-squared of model-2 will be less than that of model-1 The r-squared of model-2 increases, but the complexity of model-2 also increases The r-squared of model-2 decreases, but the complexity of model-2 also increases Nothing can be said about the r-squared of model-2

49 50 51

52 53 54

51. C2 Clustering

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30



31 32 33

34 35 36

37 38 39

✓ Both the statements are correct

40 41 42

43 44 45

46 47 48

49 50 51

51

52 53 54

Consider the following two Statements:

Statement 1: The distance between 2 clusters is the maximum distance between any 2 points in the clusters in complete linkage.Statement 2: Most of the time Complete linkage will produce unstructured dendograms.

Answer Options

Select any one option

Clear Ans

 Statement 1 is correct and Statement 2 is wrong Statement 1 is wrong and Statement 2 is correct ✓ Both the statements are correct Both the statements are wrong

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

Answer Options

Select any one option

Clear Ans

 3 2 1 0

49 50 51

52 53 54

174 min left

49. C3 Multiple Correct Answer

< Previous

MCQs

1 2 3
4 5 6
7 8 9
10 11 12
13 14 15
16 17 18
19 20 21

22 23 24
25 26 27
28 29 30

31 32 33
34 35 36
37 38 39

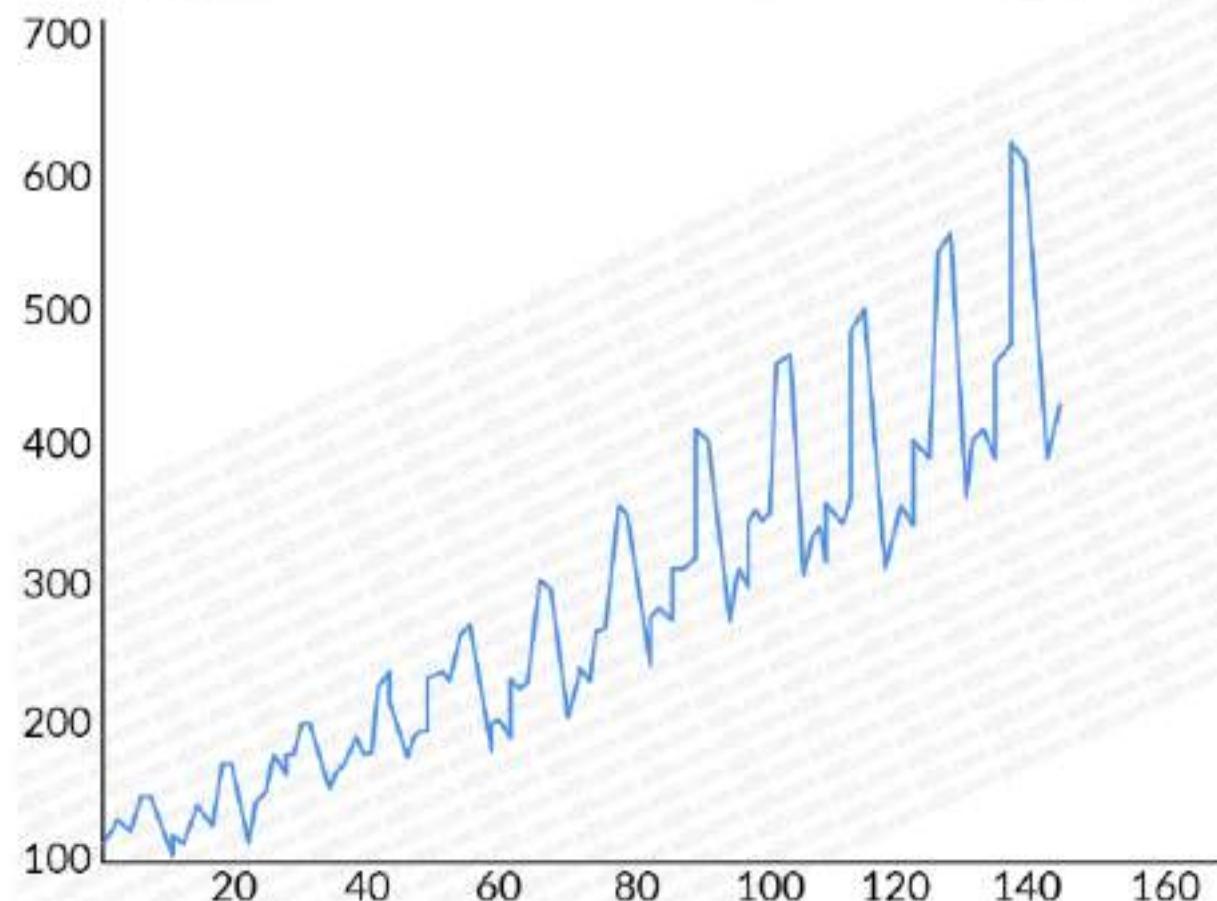
40 41 42
43 44 45
46 47 48

49 50 51
52 53 54
55 56 57

58 59 60
61 62 63
64 65 66

67 68 69
70

Refer to the image given below and choose all the statements that are correct for the given time series.(Note: Multiple options might be correct)



Answer Options

Select one or more options

Clear

 The time series has a seasonal effect. The time series has a cyclic effect. The time series has a trend effect.

Coding

48. C3 Multiple Correct Answer

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

 A: 15.5

31 32 33

 B: -1.25

34 35 36

 B: 1.75

37 38 39

 A: 16.5

40 41 42

43 44 45

46 47 48

49 50 51

52 53 54

Year	Forecast	Error	Actual
1990			15
1991			16
1992	A		13
1994		B	16
1995			14
1996			16

Answer Options

Select one or more options.

Clear Ans

 A: 15.5 B: -1.25 B: 1.75 A: 16.5

175 min left

47. C2 linear Regression

[Previous](#)[Next](#)

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

VIF (Variance Inflation Factor) is used to detect Multicollinearity. Which of the following statements is NOT true for VIF?

19 20 21

Answer Options

Select any one option

[Clear Ans](#)

22 23 24

 The VIF has a lower bound of 0

25 26 27

 The VIF has no upper bound

28 29 30



31 32 33

 VIF for a variable generally changes if you drop one of the predictor variables

34 35 36

37 38 39

 If a variable is a product of two other variables, it can have a high VIF

40 41 42

43 44 45

46 47 48

49 50 51

52 53 54

175 min left

46. C2 Logistic Regression

[Previous](#)[Next](#)

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

Answer Options

Select any one option

[Clear All](#) 0.35 0.40 0.45 0.50

46 47 48

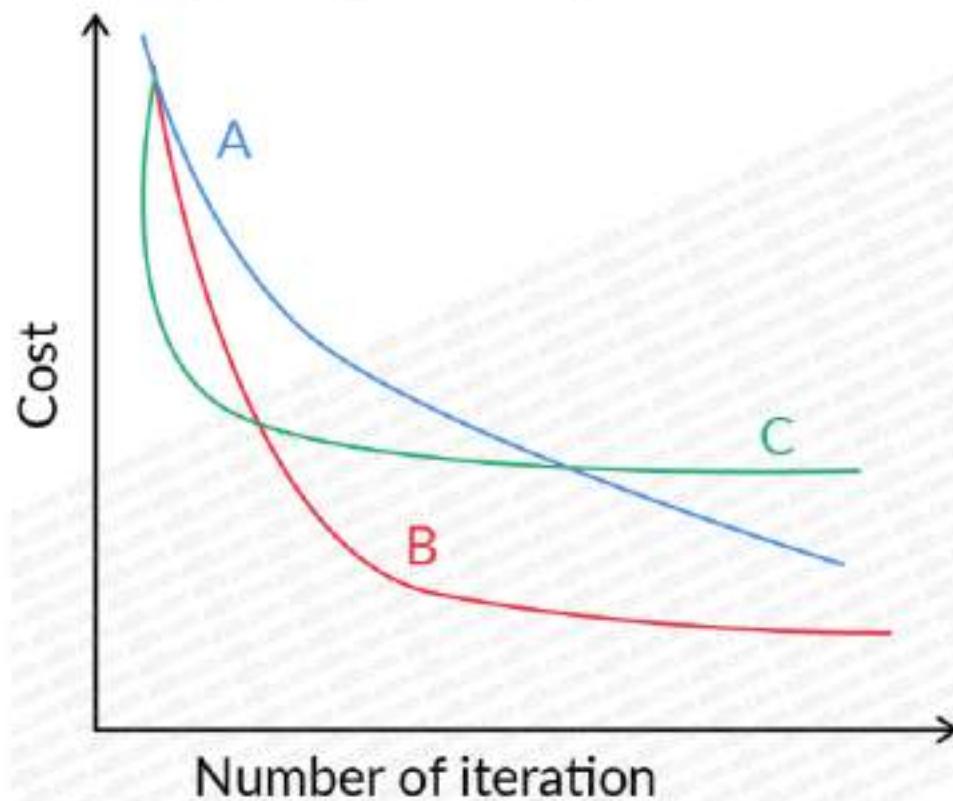
49 50 51

52 53 54

MCQs

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8
- 9
- 10
- 11
- 12
- 13
- 14
- 15
- 16
- 17
- 18
- 19
- 20
- 21
- 22
- 23
- 24
- 25
- 26
- 27
- 28
- 29
- 30
- 31
- 32
- 33
- 34
- 35
- 36
- 37
- 38
- 39
- 40
- 41
- 42

Observe the following cost function graph with different learning rates.



Which of the following statements are true about the learning rate? (More than one option may be correct.)

Answer Options

Select one or more options

Clear

The learning rate of curve C is highest among all curves

The learning rate for curve B is lower than A

The learning rate for curve B is higher than A

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30



31 32 33

34 35 36

37 38 39

40 41 42

43 44 45

46 47 48

49 50 51

52 53 54

Consider the following confusion matrix.

Total=500	Actual Positive	Actual Negative
Predicted Positive	196	20
Predicted Negative	28	256

Which among the following is the highest for the given confusion matrix?

Answer Options

Select any one option

Clear Ans

 Accuracy Precision Sensitivity Specificity

175 min left

43. C2 Logistic Regression

[Previous](#)[Next](#)

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

Recall the telecom churn example. If the log odds for churn are equal to $1/3$ for a customer, then that means -

19 20 21

Answer Options

Select any one option

[Clear Ans](#) The probability of the customer not churning is 3 times the probability of the customer churning The probability of the customer churning is 3 times more than the probability of the customer not churning The probability of the customer not churning is 4 times the probability of the customer churning The probability of the customer churning is 4 times more than the probability of the customer not churning

43 44 45

46 47 48

49 50 51

52 53 54

MCQs

Recall the telecom churn example. If the log odds for churn are equal to 1/3 for a customer, then that means -

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

Answer Options

19 20 21

 Select any one option

22 23 24

 The probability of the customer not churning is 3 times the probability of the customer churning

25 26 27

 The probability of the customer churning is 3 times more than the probability of the customer not churning
No face detected
You must not leave the frame of the camera for the duration of the test

All violations will be recorded and visible in your report

31 32 33

 The probability of the customer not churning is 4 times the probability of the customer churning

37 38 39

 The probability of the customer churning is 4 times more than the probability of the customer not churning

40 41 42

43 44 45

46 47 48

49 50 51

52 53 54

⚠ Proctoring violations detected

No face detected

You must not leave the frame of the camera for the duration of the test

Continue test

175 min left

42. C3 Time Series Forecasting

[Previous](#)[Next](#)

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30



31 32 33

34 35 36

37 38 39

40 41 42

42

43 44 45

46 47 48

49 50 51

52 53 54

Refer to the table given below and find the value of B using simple exponential smoothing with alpha = 0.5.

Year	Actual	Forecast
1990	12	14
1991	15	13
1992	13	13
1994	16	14
1995	20	A
1996	16	B

Answer Options

Select any one option

[Clear All](#) B: 15 B: 16 B: 17.5 B: 15.5

175 min left

41. C3 Advanced ML

[Previous](#)[Next](#)

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30



31 32 33

34 35 36

37 38 39

40 41 42

43 44 45

46 47 48

49 50 51

52 53 54

If you are building a model to detect if a person has a specific disease based on their vitals, which of the following algorithms would help the medical practitioner the most? (Medical practitioners prefer models with high interpretability)

Answer Options

Select any one option

[Clear All](#) Neural network Decision trees Random Forest All of the above are equally preferred

175 min left

40. C3 Time Series Forecasting

[Previous](#)[Next](#)

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

Which of the following statements regarding a stationary time series is false?

19 20 21

Select any one option

[Clear Ans](#)

22 23 24

 The p-value for the KPSS tests is 0.85

25 26 27

 The time plot will roughly have a horizontal trend with constant variance.

28 29 30



31 32 33

 The p-value for the ADF test is 0.85.

34 35 36

37 38 39

 The statistical properties of a stationary time series will be the same throughout the series, irrespective of the time window at which you observe them.

40 41 42

43 44 45

46 47 48

49 50 51

52 53 54

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

Answer Options

Select any one option

Clear Ans

- Multicollinearity is a problem when your only goal is to predict the independent variable from a set of dependent variables
- Multicollinearity is a problem when your goal is to infer the effect on the dependent variable due to independent variables
- Multicollinearity is not a problem if a variable is not collinear with your variable of interest
- Multicollinearity is not a problem if there are multiple dummy (binary) variables that represent a categorical variable with three or more categories



39

40 41 42

43 44 45

46 47 48

49 50 51

52 53 54

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30



31 32 33

34 35 36

37 38 39

40 41 42

43 44 45

46 47 48

49 50 51

52 53 54

A client approached you with a problem statement. You decided to build a multiple linear regression model on the dataset provided. The dataset consists of 40 features. Obviously all features will not be significant. Selecting the relevant features manually will be a tougher task. You can use RFE to select relevant features. RFE is an automated feature selection technique. Initially, you assumed 25 features can explain your whole data.

Which of the following commands correctly calls the RFE technique in Python? (Here "lm" is the fitted instance of multiple linear regression model)

Answer Options

Select any one option

Clear Ans

from statsmodel.feature_selection import RFE
rfe = RFE(lm, 25)
rfe = rfe.fit(X_train, y_train)

from sklearn.feature_selection import RFE
rfe = RFE(lm, 25)
rfe = rfe.predict(X_train, y_train)

from sklearn.feature_selection import RFE
rfe = RFE(lm, 25)
rfe = rfe.fit(X_train, y_train)

from RFE import feature_selection
rfe = RFE(lm, 25)
rfe = rfe.predict(X_train, y_train)

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

31 32 33

34 35 36

37 38 39

40 41 42

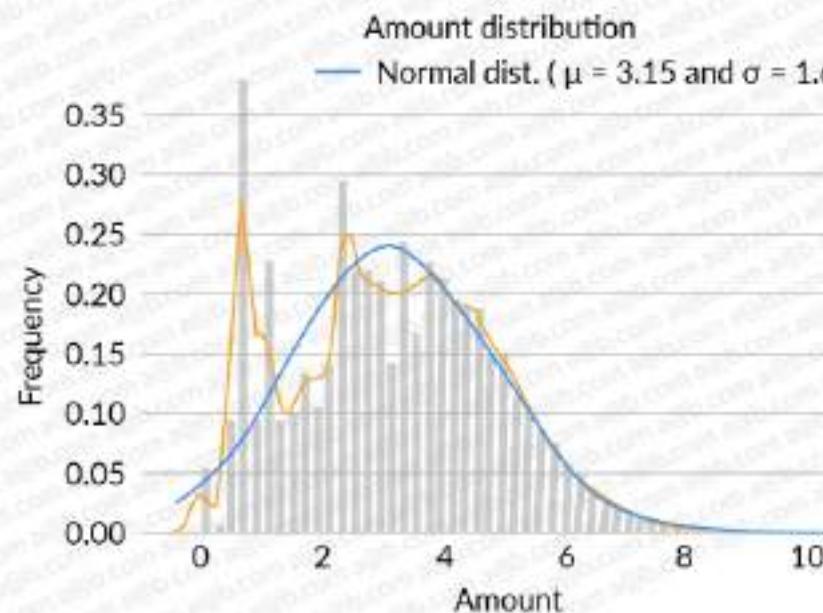
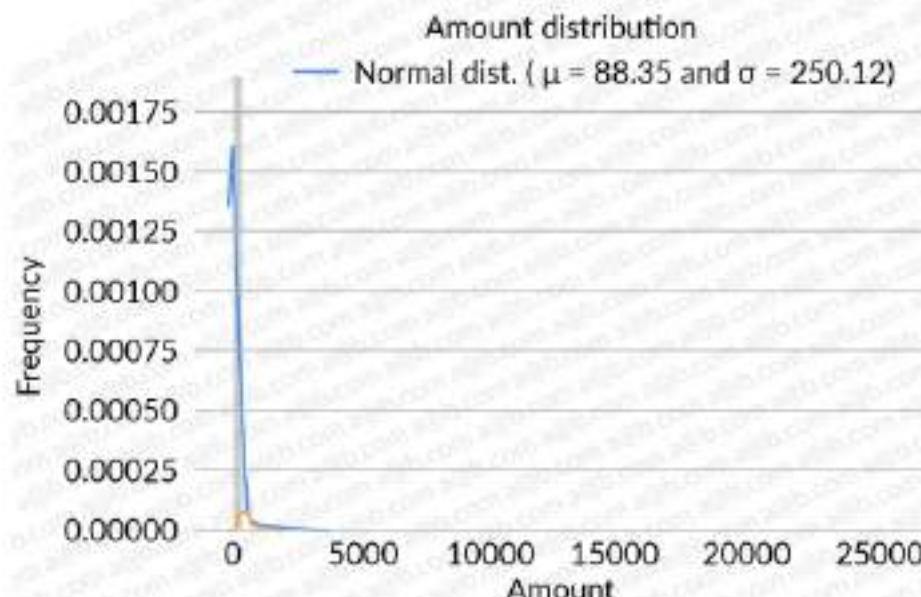
43 44 45

46 47 48

49 50 51

52 53 54

Refer to the plot below which consists of two plots A and B. Plot A represents the non transformed column "Amount" while column B represents the transformed variable "Amount". Observe the plot B carefully and identify which kind of the transformation is applied to the variable "Amount"?



Answer Options

Select any one option

Clear Ans

Standardisation

Log Transformation

Min-Max Scaling

175 min left

36. C3 Advanced ML

[Previous](#)[Next](#)

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

The primary issue with algorithms having high variance is

Answer Options

Select any one option

[Clear Ans](#)

- Model produced is too sensitive to the input data – it will also learn what shouldn't be learned from the data given.
- Model becomes too complex to deal with.
- Takes a lot of time and computational effort to train such models.
- All of the above



36

37 38 39

40 41 42

43 44 45

46 47 48

49 50 51

52 53 54

35. C3 Multiple Correct Answer.

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

Consider the following confusion matrix that summarises the prediction made by a model.

		Predicted	
		Achieved = YES	Achieved = No
Actual	Achieved = YES	100(TP)	40(FN)
	Achieved = No	50(FP)	60(TN)

Which of the following option(s) is/are correct?

Answer Options

Select one or more options

Clear All

Accuracy is ~0.72



Precision is ~0.67

Recall is ~0.7

F1 score is ~0.69

35. C3 Multiple Correct Answer.

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

<<

31 32 33

34 35 36

37 38 39

40 41 42

43 44 45

46 47 48

49 50 51

52 53 54

Consider the following confusion matrix that summarises the prediction made by a model.

		Predicted	
		Achieved = YES	Achieved = No
Actual	Achieved = YES	100(TP)	40(FN)
	Achieved = No	50(FP)	60(TN)

Which of the following option(s) is/are correct?

Answer Options

Select one or more options

Clear All

Accuracy is ~0.72

Precision is ~0.67

Recall is ~0.7

F1 score is ~0.69

176 min left

34. C2 Clustering

[Previous](#)[Next](#)

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

Clustering is used to identify the below?

Answer Options

Select any one option

[Clear All](#) Data distribution Correlation among the data points Principal components Subgroups in the data

34 35 36

37 38 39

40 41 42

43 44 45

46 47 48

49 50 51

52 53 54

176 min left

33. C2 Clustering

< Previous

Next

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

31 32 33

34 35 36

37 38 39

40 41 42

43 44 45

46 47 48

49 50 51

52 53 54

In hierarchical clustering, the shortest distance and the maximum distance between points in two clusters are defined as _____ and _____ respectively.

Answer Options

Select any one option

Clear Ans

 Single linkage and Complete linkage Complete linkage and Single linkage Single linkage and Average linkage Complete linkage and Average linkage

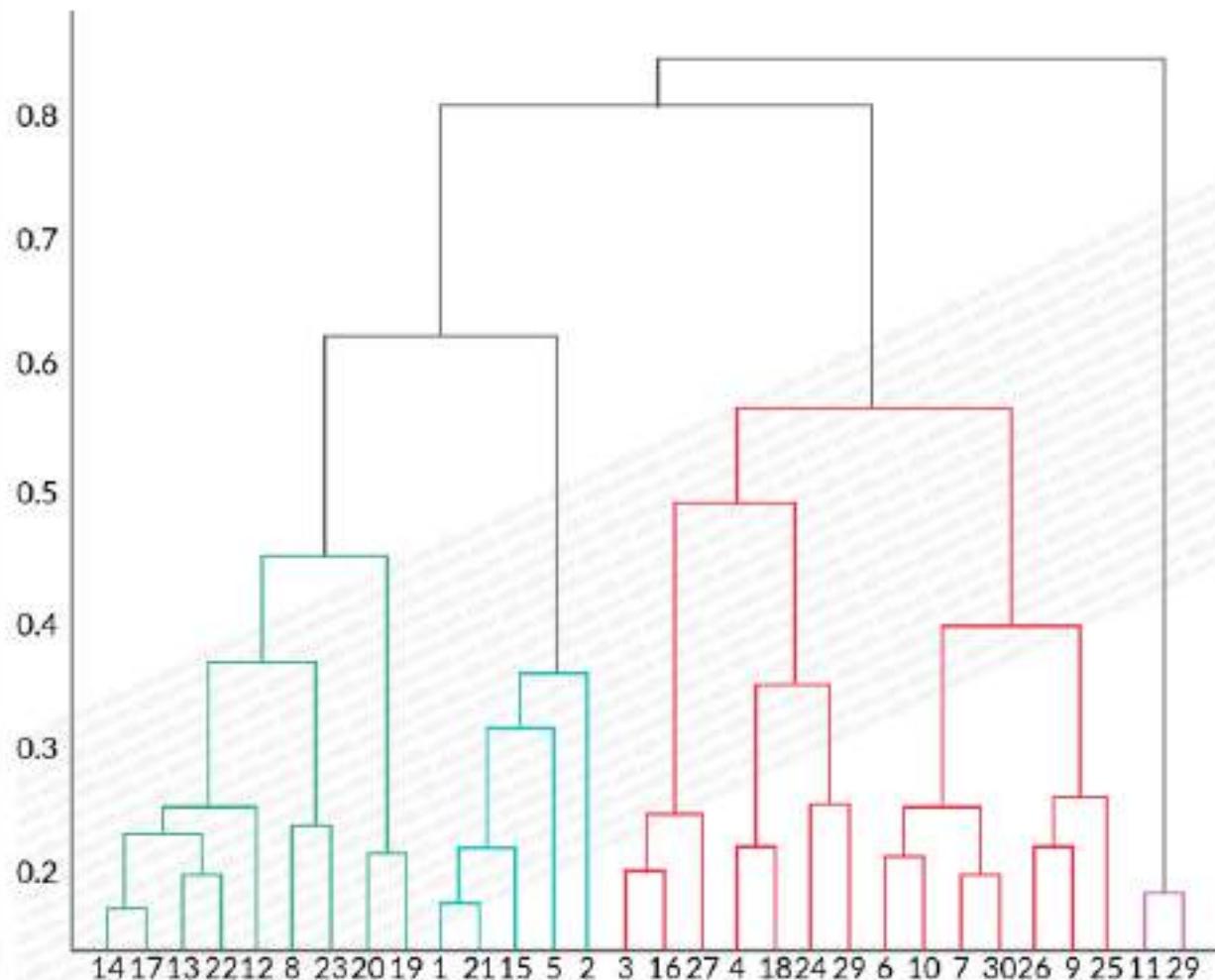
Score:

1	1	3
4	5	6
7	8	9
10	11	12
13	14	10
16	17	18
19	20	21
22	23	24
25	26	27
28	29	30
31	32	
34	35	36
37	38	
39	39	39
41	42	
43	43	
47	48	
50	51	
53	54	
55	57	
58	60	
61	63	
64	66	
67	69	
70		

Coding:

1	2	3
4		

You obtained the following dendrogram after performing K-means clustering on a dataset. Which of the following statements can be concluded from the dendrogram?



Answer Options:

Select any one option:

- The initial number of clusters is 6.
- There are 25 data points used in the above clustering algorithm.
- Single linkage is used to define the distance between two clusters in the above dendrogram.
- The elbow dendrogram in tecplot.com is not suitable for K-Means clustering.

176 min left

31. C2 linear Regression

[Previous](#)[Next](#)

MCQs:

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

If the coefficient of determination is 0.47 between a dependent variable and an independent variable, This denotes that:

22 23 24

Select any one option.

[Clear A](#)

25 26 27

 The relationship between the two variables is not strong.

28 29 30

 The correlation coefficient between the two variables is also 0.47

31 32 33

 47% of the variance in the independent variable is explained by the dependent variable

34 35 36

 47% of the variance in the dependent variable is explained by the independent variable.

37 38 39

40 41 42

43 44 45

46 47 48

49 50 51

52 53 54

55 56 57

58 59 60

61 62 63

MCQs:

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

31 32 33

34 35 36

37 38 39

40 41 42

43 44 45

46 47 48

49 50 51

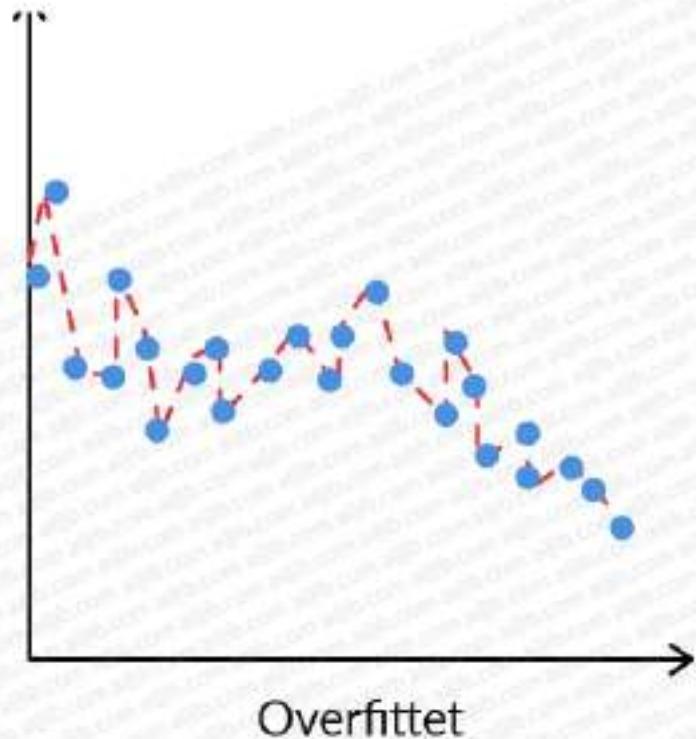
52 53 54

55 56 57

58 59 60

61 62 63

You are given an overfitted model as shown in the image given below. Which of the following options would reduce the complexity of the model, so that it aligns with the principle of bias-variance trade-off?



Answer Options

Select any one option

[Clear A](#)

- Increasing the number of datapoints in the training set will help fulfil this requirement.
- If the model is using a random forest regressor, then the number of trees in the forest needs to be increased.
- If the model is using a decision tree regressor, then the depth of the tree needs to be increased.

MCQs:

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

31 32 33

34 35 36

37 38 39

40 41 42

43 44 45

46 47 48

49 50 51

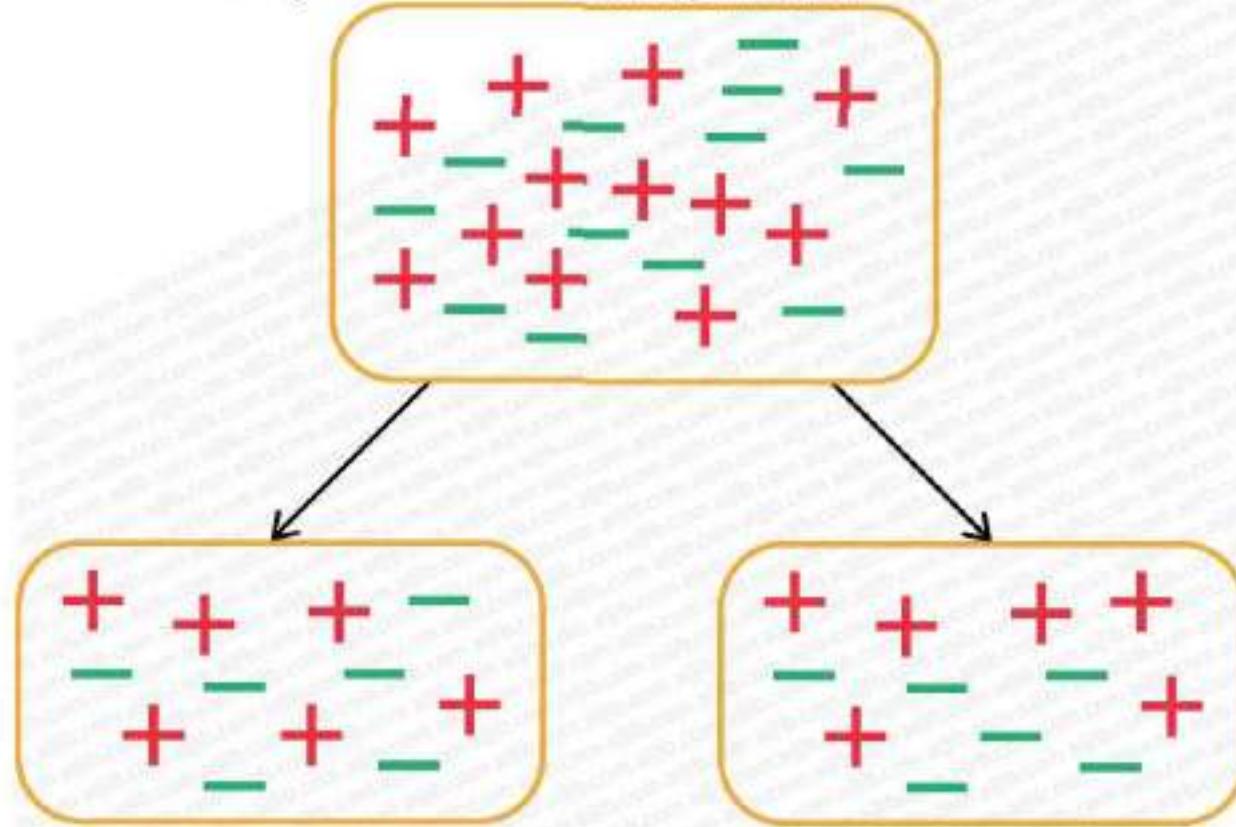
52 53 54

55 56 57

58 59 60

61 62 63

Refer to the decision tree given below and choose what is wrong with this tree.



Answer Options

Select any one option

Clear A

- There is nothing wrong with this decision tree.
- The split is done incorrectly. The leaf nodes are as impure as the root node.
- The tree given above is not a decision tree as both the leaf nodes are heterogeneous.

177 min left

28. C2 Clustering

[Previous](#)[Next](#)

MCQs:

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

31 32 33

34 35 36

37 38 39

40 41 42

43 44 45

46 47 48

49 50 51

52 53 54

55 56 57

58 59 60

61 62 63

A client has approached you for a problem statement that requires the use of clustering. You decided to model the problem statement with hierarchical clustering. Consider datasets having 'n' data points.

Which of the following statements is true for the above problem statement?

Answer Options

Select any one option:

[Clear A](#)

- 'n*n' distance matrix should be calculated for the mentioned problem statement
- Initially 'n' clusters are formed for the mentioned problem statement
- The output for the problem statement above is a dendrogram
- All the above

177 min left

27. C2 Clustering

[Previous](#)[Next](#)

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

31 32 33

34 35 36

37 38 39

40 41 42

43 44 45

46 47 48

49 50 51

52 53 54

55 56 57

58 59 60

61 62 63

Which of the following is not true for Hopkins statistic?

Answer Options

Select any one option.

[Clear A](#)

- Hopkins statistic decides if the data is suitable for clustering or not
- Hopkins statistic lie between -1 and 1
- If the Hopkins statistic comes out to be 0, then the data is uniformly distributed
- If the Hopkins statistic comes out to be 1, then the data highly suitable for clustering

INFO

Which of the following is not true for Hopkins statistic?

- 1. 0 < H < 1
- 2. H = 0
- 3. H = 1
- 4. H = 11
- 5. H = 14
- 6. H = 17
- 7. H = 21

Answer Options

- 8. H = 24

Difficulty level

- 9. 25

Required statistics: deviation of the data is suitable for clustering

- 10. 29

Hopkins statistic is between -1 and 1

- 11. 33

If the Hopkins statistic comes out to be 0, then the data is uniformly distributed

- 12. 41

If the Hopkins statistic comes out to be 1, then the data is highly clustered

- 13. 47

If the Hopkins statistic comes out to be 0.5, then the data is moderately clustered

- 14. 52

If the Hopkins statistic comes out to be 1, then the data is highly clustered

- 15. 57

If the Hopkins statistic comes out to be 0.5, then the data is moderately clustered

- 16. 62

If the Hopkins statistic comes out to be 0, then the data is uniformly distributed

- 17. 67

If the Hopkins statistic comes out to be 1, then the data is highly clustered

 **Proctoring violations detected**

No face detected

You must not leave the frame of the camera for the duration of the test

All violations will be recorded and visible in your report

Continue test

177 min left

26. C3 Time Series Forecasting

< Previous

Next >

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

31 32 33

34 35 36

37 38 39

40 41 42

43 44 45

46 47 48

49 50 51

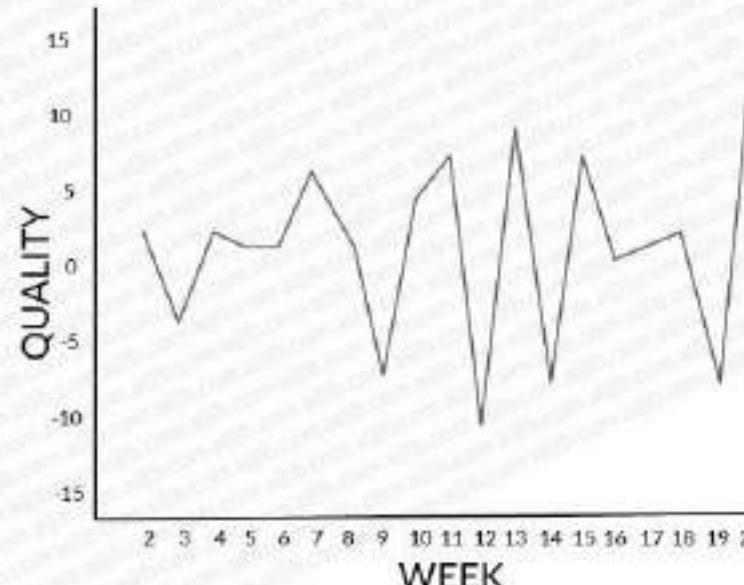
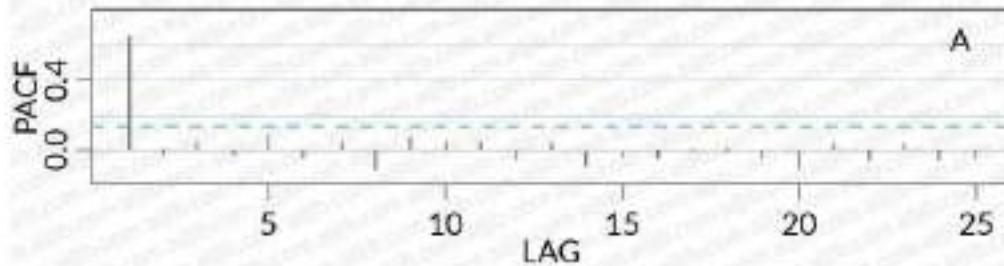
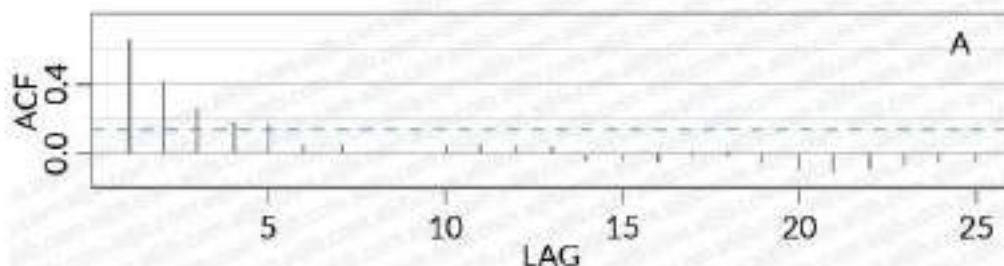
52 53 54

55 56 57

58 59 60

61 62 63

Refer to the image with 3 sub-images A, B and C. Find the order (p,d,q) of the time series for the ARIMA process. Image C is the differenced time series of the original series. The original series has the data from week-1 to week-20.



Answer Options

Select any one option

Clear All

 (1,0,1) (1,1,5) (1,2,5)

177 min left

25. C2 Logistic Regression

[Previous](#)[Next](#)

MCQs:

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

31 32 33

34 35 36

37 38 39

40 41 42

43 44 45

46 47 48

49 50 51

52 53 54

55 56 57

58 59 60

61 62 63

For a completely random binary classification model, what will be the area under the curve of the ROC graph?

Answer Options

Select any one option.

[Clear A](#) 0 0.25 0.5 1

177 min left

24. C3 Advanced ML

< Previous

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

31 32 33

34 35 36

37 38 39

40 41 42

43 44 45

46 47 48

49 50 51

52 53 54

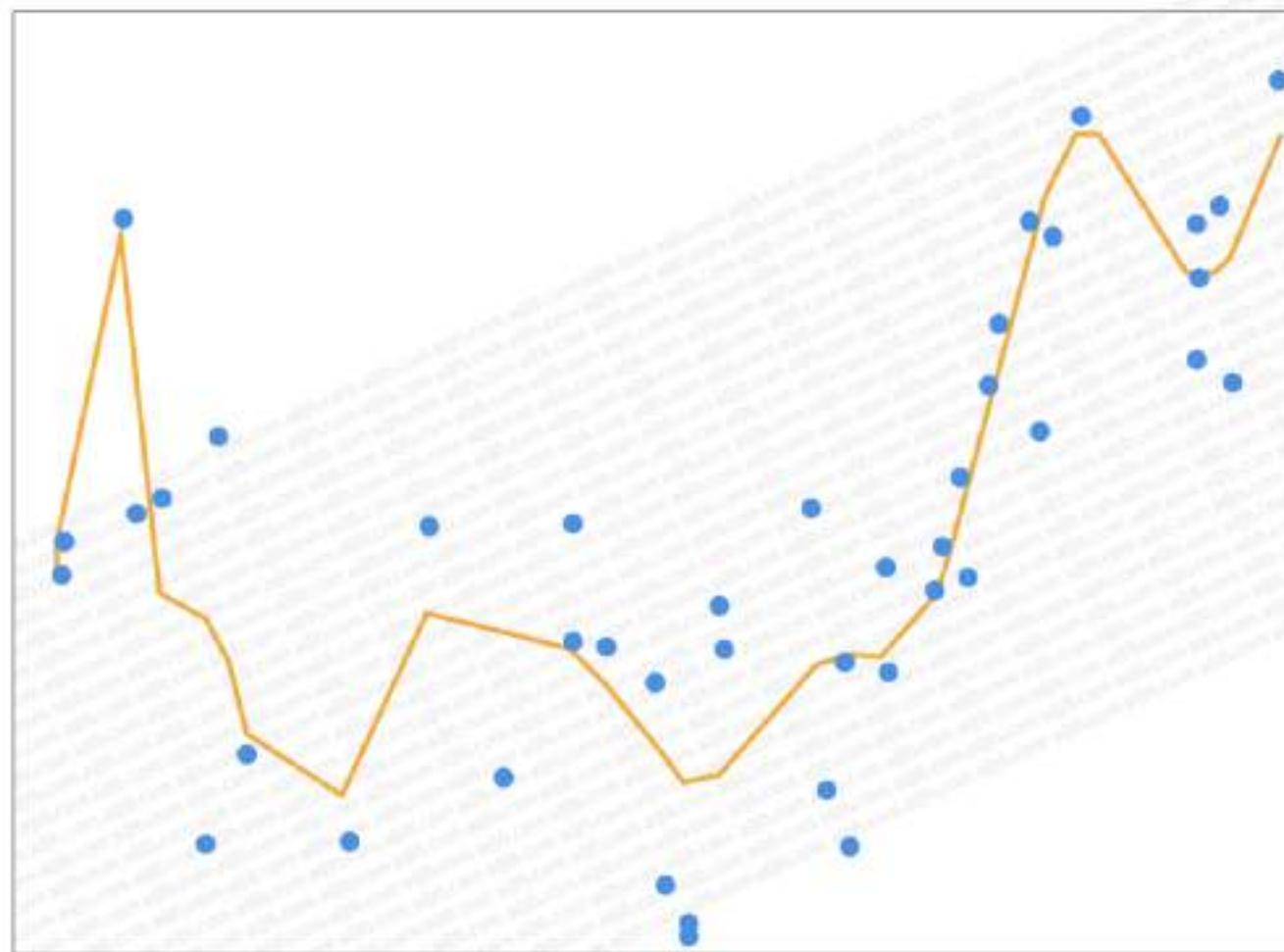
56 56 56

61 62 63

64 65 66

67 68 69

70



Answer Options

Select any one option

 He will get an underfitting model as the increase in the polynomial degree will not help reduce the model bias. He will get an overfitting model as the current model also overfits the train data.

179 min left

8. C3 Multiple Correct Answer

[Previous](#)[Next](#)

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

 [Answer Options](#)

25 26 27

Select one or more options

[Clear Answer](#)

28 29 30

 Accuracy is ~0.72

31 32 33

34 35 36

37 38 39

Submit Test

 Recall is ~0.7 F1 score is ~0.69

Consider the following confusion matrix that summarises the prediction made by a model.

		Predicted	
		Achieved = YES	Achieved = No
Actual	Achieved = YES	100(TP)	40(FN)
	Achieved = No	50(FP)	60(TN)

Which of the following option(s) is/are correct?

179 min left

7. C2 Logistic Regression

[Previous](#)[Next](#)

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

31 32 33

34 35 36



Recall the telecom churn example. If the log odds for churn are equal to 0 for a customer, then that means -

- Select any one option [Clear Answer](#)
- There is no chance of the customer churning
- The probability of the customer churning is equal to the probability of the customer not churning
- The probability of the customer churning is very small compared to the probability of the customer not churning
- The probability of the customer churning is very large compared to the probability of the customer not churning

179 min left

6. C3 Advanced ML

[Previous](#)[Next](#)

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

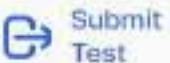
22 23 24

25 26 27

28 29 30

31 32 33

34 35 36



Suppose you built a random forest classifier model. You observed that the accuracy of the model turns out to be 99% on the train data but performs badly on test data.

The senior data scientist of your company suggested you to create a model on the subsets of the training data and test the same model on different subsets of the training data.

Which of the following statements is NOT true considering the above scenario?

Answer Options

Select any one option

[Clear Answer](#)

The senior data scientist is talking about the Cross-validation scheme here.

Cross-validation schemes can only give you reliable insights if your data is very large.

OOB score is similar to cross-validation score in random forest.

Cross-validation score is likely to be less than the accuracy of the model you built.

179 min left

5. C3 Advanced ML

[Previous](#)[Next](#)

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

31 32 33

34 35 36

[Submit Test](#)

Decision Tree is a high variance model. Which of the following correctly explains this statement?

Answer Options

Select any one option

[Clear Answer](#)

The number of attributes on both the sides of the decision tree is not the same.

The decision tree building process is a top-down approach.

The entire structure of the tree changes with small variations in the input data.

All the above

Score

1

2

3

4

5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

32

33

34

35

36

37

38

39

40

41

42

43

44

45

46

47

48

49

50

51

52

53

54

55

56

57

58

59

60

61

62

63

64

65

66

67

68

69

70

71

72

73

74

75

76

77

78

79

80

81

82

83

84

85

86

87

88

89

90

91

92

93

94

95

96

97

98

99

100

101

102

103

104

105

106

107

108

109

110

111

112

113

114

115

116

117

118

119

120

121

122

123

124

125

126

127

128

129

130

131

132

133

134

135

136

137

138

139

140

141

142

143

144

145

146

147

148

149

150

151

152

153

154

155

156

157

158

159

160

161

162

163

164

165

166

167

168

169

170

171

172

173

174

175

176

177

178

179

180

181

182

183

184

185

186

187

188

189

190

191

192

193

194

195

196

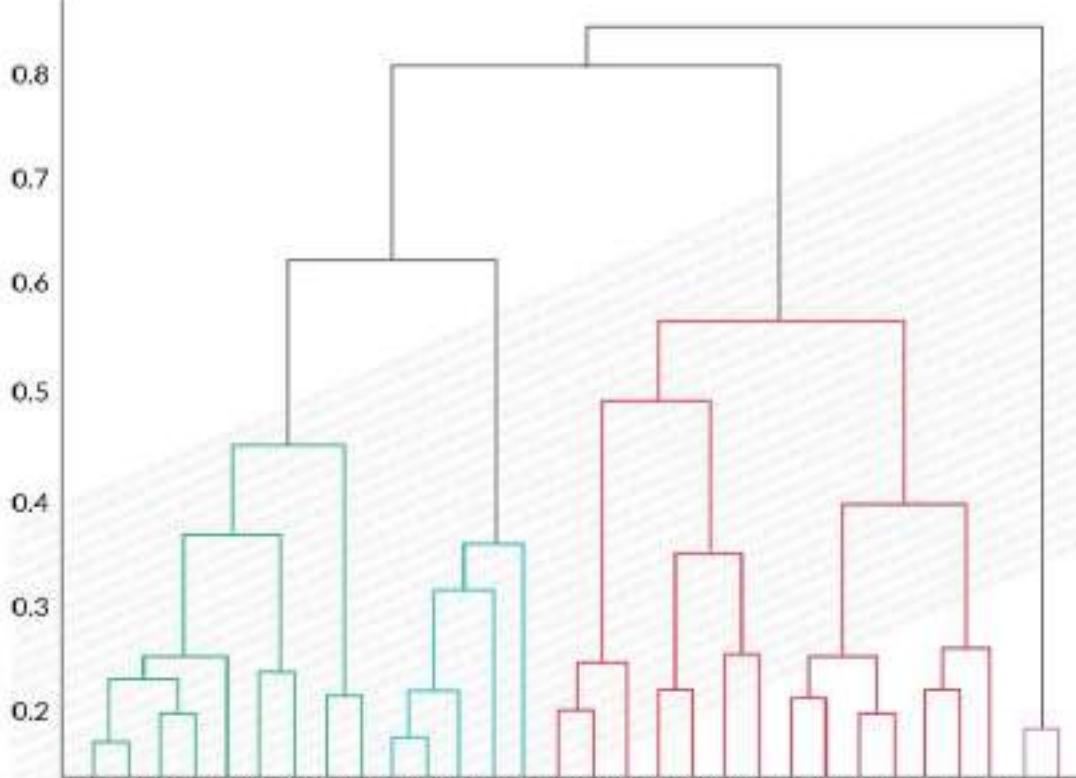
197

198

199

200

You observed the following dendrogram after performing k-means clustering on a dataset. Which of the following statements can be concluded from the dendrogram?

**Answer Options**

- The total number of clusters is 6.
- There are 23 data points used in the above clustering algorithm.
- Single linkage is used to define the distance between two clusters in the above dendrogram.
- The above dendrogram interpretation is not possible for K-Means clustering.

[Close Answer](#)

3. C2 linear Regression

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

31 32 33

34 35 36

37 38 39

40 41 42

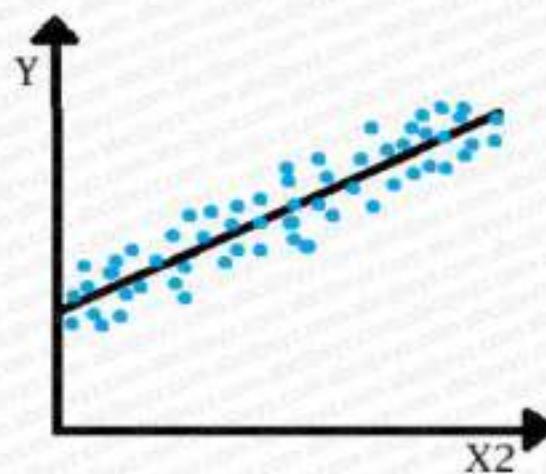
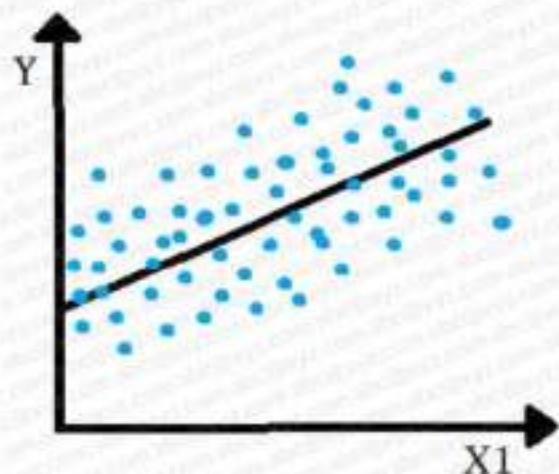
43 44 45

46 47 48

49 50 51

52 53 54

For the same dependent variable Y, two models were created using the independent variables X1 and X2. The following two graphs represent the fitted line on the scatterplot. (Both of the graphs are on the same scale)



Which of the following is true about the residuals in these two models?

Answer Options

Select any one option

Clear Answer

- The sum of residuals in model 2 is higher than model 1
- The sum of residuals in model 1 is higher than model 2
- Both have the same sum of residuals
- Nothing can be said about the sum of residuals from the graph



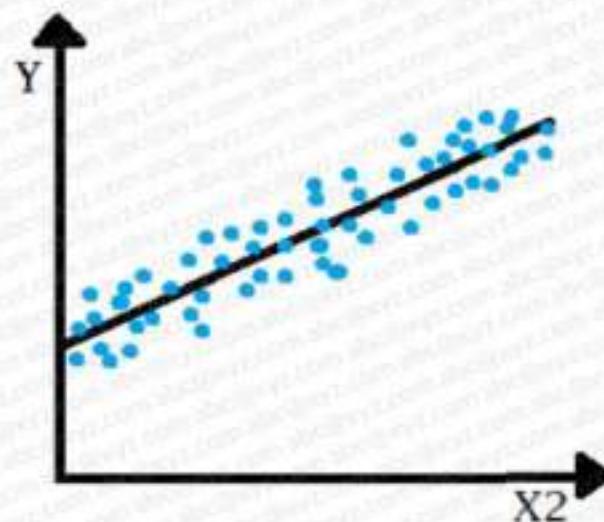
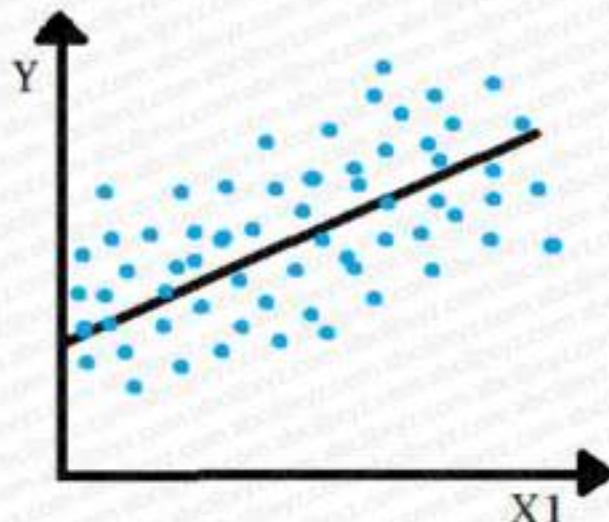
Submit Test

3. C2 Linear Regression

MCQs

- 1 2 3
4 5 6
7 8 9
10 11 12
13 14 15
16 17 18
19 20 21
22 23 24
25 26 27 <<
28 29 30
31 32 33
34 35 36
37 38 39
40 41 42
43 44 45

For the same dependent variable Y, two models were created using the independent variables X1 and X2. The following two graphs represent the fitted line on the scatterplot. (Both of the graphs are on the same scale)



Which of the following is true about the residuals in these two models?

Answer Options

Select any one option

[Clear Answer](#)

- The sum of residuals in model 2 is higher than model 1
- The sum of residuals in model 1 is higher than model 2
- Both have the same sum of residuals



Submit Test

180 min left

2. C3 Time Series Forecasting

[Previous](#)[Next >](#)

MCQs

1 **2** 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

Select any one option

[Clear Answer](#) Simple moving average SARIMA Holt-Winters' SARIMAX

30 31 32

33 34 35

36 37 38

39 40 41

42 43 44

45 46 47



180 min left

1. C3 Time Series Forecasting

[Previous](#) [Next](#)

MCG-8

1 2 3

4 5 6

7 8 9

19 11 12

13 14 15

16 17 18

10 20 30

22 23 24

Consider the following two problem statements:-

Problem statement A: You have to predict the price of a stock for the next day based on the stock prices of the previous four days.

Problem statement B: You have to predict the price of a stock for the next month based on monthly data of the last 5 years. The stock price has an increasing trend.

Which of these forecasting methods are correctly mapped with the problem statements?

Answer Options

Select any one option

Clear Answer

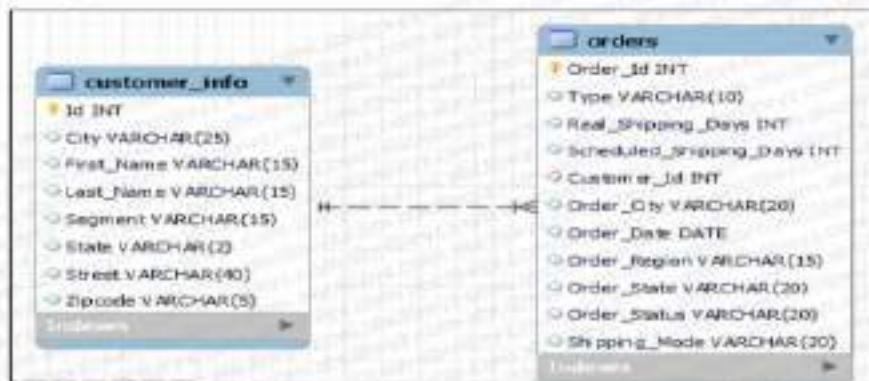
- A: Simple moving average, B: Holt's Method
 - A: Naïve method, B: Holt-Winters' method
 - A: Simple moving average, B: Naïve method



13 14 15
 16 17 18
 19 20 21
 22 23 24
 25 26 27
 28 29 30
 31 32 33
 34 35 36
 37 38 39
 40 41 42
 43 44 45
 46 47 48
 49 50 51
 52 53 54
 55 56 57
 58 59 60
 61 62 63
 64 65 66
 67 68 69
 70

INPUT TABLE

You are given two tables - customer_info and orders whose schema is given below:

**QUERY:**

Display all the details (all columns) from the **orders** table
where

- + State is AZ and
- + Street contains the word 'Silver'

Order them by Order_Id in ascending order.

Note – The Id column in customer_info is common with the Customer_Id column in the orders table.

OUTPUT COLUMNS:

All column of orders table.

Note that the coding console automatically converts the casing of the columns to upper case.

PLEASE NOTE:

- Use the original column names only. Any other aliases or column names would lead to an error.
- Keep the sequence of the columns same as in the original table.

Coding

1 2 3

4



Submit Test

Schema

Table structure

department

Name	Type	Description
Id	int	
Name	varchar(20)	

category

174 min left

3. GroupBy

< Previous Next >

13 14 15
 16 17 18
 19 20 21
 22 23 24
 25 26 27
 28 29 30
 31 32 33
 34 35 36
 37 38 39
 40 41 42
 43 44 45

INPUT TABLE

You're given a **orders** table and the columns in the orders table are shown below:

orders		
Order_Id	INT	
Type	VARCHAR(10)	
Real_Shipping_Days	INT	
Scheduled_Shipping_Days	INT	
Customer_Id	INT	
Order_City	VARCHAR(20)	
Order_Date	DATE	
Order_Region	VARCHAR(15)	
Order_State	VARCHAR(20)	
Order_Status	VARCHAR(20)	
Shipping_Mode	VARCHAR(20)	
Count(orders)		

QUERY

- Calculate count of all the orders.
- Where Order_State is Maharashtra:
 - Note - Use the alias of **oc** for count of orders.
- Group the results by **Type**.
- Order them by **oc** in **ascending** order.

OUTPUT COLUMNS

oc, Type

Here's an image showing how a sample output would look like. :

OC	TYPE
45	CASH
89	PAYMENT
108	TRANSFER
151	DEBIT

Note that the coding console automatically converts the casing of the columns to uppercase.

PLEASE NOTE:

- Use the alias of 'oc' for the count of orders. Any other alias or column name would lead to an error.
- Please keep the sequence same as mentioned in the output columns. For example in the above scenario, if you display **Type,oc** instead of **oc,Type** you'll get an error.



Submit Test

Schema

174 min left

2. GroupBy and OrderBy

[Previous](#) [Next](#)

13 14 15
 16 17 18
 19 20 21
 22 23 24
 25 26 27
 28 29 30
 31 32 33
 34 35 36
 37 38 39
 40 41 42
 43 44 45
 46 47 48
 49 50 51 <<
 52 53 54
 55 56 57
 58 59 60
 61 62 63
 64 65 66
 67 68 69
 70

Coding

1 2 3
 4



INPUT TABLE

You're given a `orders` table and the columns in the `orders` table are shown below:

<code>orders</code>	
Order_Id	INT
Type	VARCHAR(30)
Real_Shipping_Days	INT
Scheduled_Shipping_Days	INT
Customer_Id	INT
Order_City	VARCHAR(20)
Order_Date	DATE
Order_Region	VARCHAR(15)
Order_State	VARCHAR(20)
Order_Status	VARCHAR(20)
Shipping_Mode	VARCHAR(20)
Total Orders	

QUERY

- Calculate count of all the orders
 - where the `Order_Status` is **Gujarat**
- where the `Order_Status` is **PENDING**.
 - Note - Use the alias of `oc` for count of orders.
- Group the results by `Order_City`
- Order them by `oc` & `Order_City` in *ascending order*.

OUTPUT COLUMNS

`oc, Order_City`

Here's an image showing how a sample output would look like.:

OC	ORDER_CITY
1	Jamnagar
1	Rajkot
1	Vadodara
2	Jodhpur
3	Bhavnagar
4	Surat

Note that the coding console automatically converts the casing of the columns to upper case

PLEASE NOTE:

- Use the alias of `'oc'` for the count of orders. Any other alias or column name would lead to an error.

174 ms left

1. kth Largest

< Previous

Next >

Python 3

Problem Statement:

Given a list of integers and another integer 'K', find the k^{th} largest integer in the list. If there are less than ' k ' distinct elements in the list, then you need to output -1. (Refer to the sample inputs and outputs for details).

Note: $k=1$ means you have to find the largest element.

Input Format:

Line 1 contains a list of integers

Line 2 contains a positive integer $k > 0$

Output Format:

An integer, k^{th} largest integer.

Examples:**Sample Input 1:**

```
[2, 3, 1, 5, 6, 2, 1]
```

```
4
```

Sample Output 1:

```
2
```

Sample Input 2:

```
[2, 3, 1, 5, 6, 2, 1]
```

```
6
```

Sample Output 2:

```
-1
```

Function description

Complete the `kthLargest` function in the editor below. It has the following parameter(s):

Name	Type	Description
<code>k</code>	INTEGER	
<code>arr</code>	INTEGER ARRAY	

Return The function must return an INTEGER denoting the k^{th} largest integer in the array.

Constraints

Input format for debugging

Sample Testcases

```

1 import sys
2
3
4 def kthLargest(arr, k):
5     # Write your code here
6
7
8
9 def main():
10    import ast
11    arr = []
12    arr = ast.literal_eval(input())
13    k = int(input())
14    result = kthLargest(arr, k)
15    print(result)
16
17
18 if __name__ == "__main__":
19    main()

```

Coding

1

2

3

4



Submit Test

Console

Custom Test Case

You must ex

124 min left

70. C3- Advanced ML

[Previous](#) [Next](#)

13 54 15

16 17 19

19 20 21

22 23 24

25 26 27

28 29 30

31 32 33

34 35 36

37 38 39

40 41 42

43 44 45

46 47 48

49 50 51

52 53 54

55 56 57

58 59 60

61 62 63

64 65 66

67 68 69

70

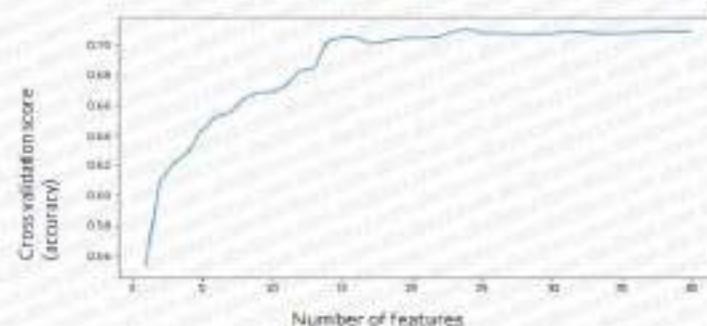
Coding

1 2 3

4



Observe the following graph for a logistic regression model. The graph determines the Cross-validation score based on the different number of features.



Which of the following statements is NOT true based on the above graph?

Answer Options[Edit/Change option](#)[Clear Answer](#)

- There are 40 features in the model.
- The performance of the model is lowest with one feature.
- The performance of the model is optimum with the number of features between 15 and 25.
- Cross-validation score can not be used in deciding the optimum number of features.

174 min left

69. C2 Multiple correct answer

[Previous](#)

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

31 32 33

34 35 36

Answer Options

37 38 39

Select one or more options

[Clear All](#)

from sklearn.linear_model import LogisticRegression
lr = LogisticRegression()
lr.fit(X_train, y_train)

import statsmodel.api as sm
lr = sm.GLM(y_train,(sm.add_constant(X_train)),
family = sm.families.Binomial())
lr.fit()

from sklearn.linear_model import LogisticRegression
lr = LogisticRegression()
lr.predict(X_train, y_train)

import statsmodel.api as sm
lr = sm.GLM(y_train,(sm.add_constant(X_train)),
family = sm.families.Binomial())
lr.predict()

Coding

1 2 3

4

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

31 32 33

34 35 36

37 38 39

40 41 42

43 44 45

46 47 48

49 50 51

52 53 54

55 56 57

58 59 60

61 62 63

64 65 66

67 68 69

70

Choose the correct option from the following.
The difference between '+' and '*' quantifier is _____.

Answer Options

Select any one option

Clear All

- '+' needs the preceding character to be present at least once whereas '*' does not need the same.
- '*' needs the character to be present at least once whereas '+' does not need the same.
- Both the quantifiers have the same functionality.
- None of the above.



Coding

1 2 3

4

174 min left

67. C2 linear Regression

[Previous](#)[Next](#)

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

31 32 33

34 35 36

37 38 39

40 41 42

43 44 45

46 47 48

49 50 51

52 53 54

55 56 57

58 59 60

61 62 63

64 65 66

67 68 69

70

Suppose you run a regression with one of the feature variables T, with all the remaining feature variables. The R-squared of this model was found out to be 0.8. What will be the VIF for variable T?

Answer Options

Select any one option

[Clear All](#) 1.56 2.77 3.33 5.00

67

Coding

1 2 3

4

174 min left

66. C3 Time Series Forecasting

[Previous](#)[Next](#)

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

31 32 33

34 35 36

37 38 39

Answer Options

40 41 42

Select any one option

[Clear All](#) 3 2 1 0

49 50 51

52 53 54

55 56 57

58 59 60

61 62 63

64 65 66

67 68 69

70

Coding

1 2 3

4

174 min left

65. C2-Decision Trees

[Previous](#)[Next](#)

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

31 32 33

34 35 36

37 38 39

40 41 42

43 44 45

46 47 48

49 50 51

52 53 54

55 56 57

58 59 60

61 62 63

64 65 66

67 68 69

70

Suppose you train a decision tree with the following data. Which feature should we split on at the root?

X	Y	Z	V
T	T	F	1
F	F	F	0
T	T	T	0
F	T	T	1

Answer Options

Select any one option

[Clear](#) X Y Z Cannot be determined

Coding

1 2 3

4

174 min left

64. C2-Basics of NLP and Text Mining

[Previous](#)[Next](#)

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

31 32 33

34 35 36

37 38 39

40 41 42

43 44 45

46 47 48

49 50 51

52 53 54

55 56 57

58 59 60

61 62 63

64 65 66

67 68 69

70

What is the Levenshtein distance between 'decade' and 'dictate'?

Answer Options

Select any one option

[Clear All](#) 3 4 5 6

Coding

1 2 3

4

174 min left

63. C2 Multiple correct answer

[Previous](#)[Next](#)

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

31 32 33

34 35 36

37 38 39

40 41 42

43 44 45

46 47 48

49 50 51

52 53 54

55 56 57

58 59 60

61 62 63

64 65 66

67 68 69

70

Which of the following is NOT a methodology by which you can identify the optimal number of clusters for K-means clustering? (More than one option may be correct)

Answer Options

Select one or more options

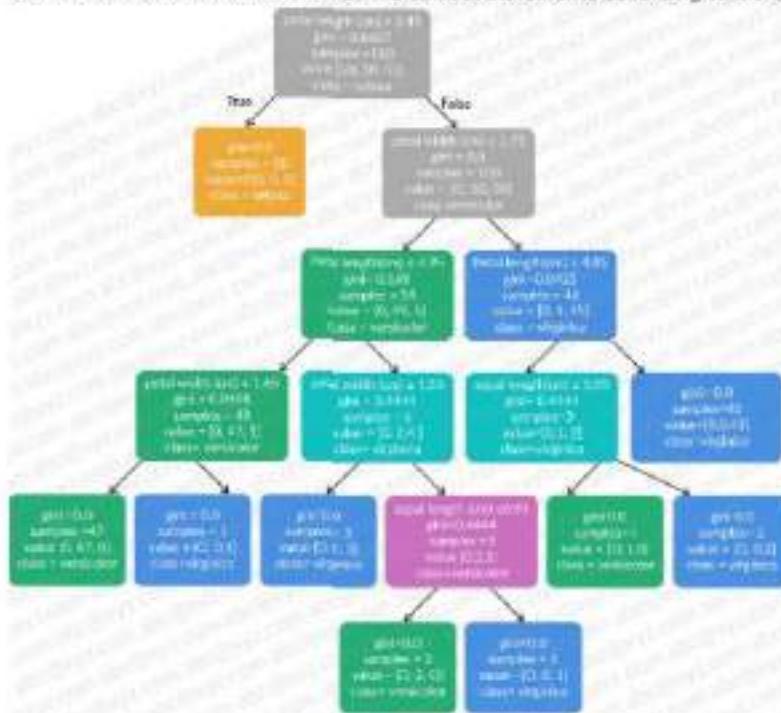
[Clear All](#) Dendrogram inspection method Elbow Method Single Linkage Method Silhouette score

Coding

1 2 3

4

Refer to the decision and choose all the correct statements according to the decision tree provided. (More than one option may be correct.)



Answer Options

Select one or more options

Class 1

- The given tree is an overfitting tree.
 - The given tree will be having a stable performance. If we change one row from the training data,

If the total length is more than 3.45 mm, it's probably likely that the flower is either a *Veronica* or *Vinca*.

- The given tree is a high variance model

174 min left

61. C3 Multiple Correct Answer

[Previous](#)[Next](#)

13. 14. 15.

16. 17. 18.

19. 20. 21.

22. 23. 24.

25. 26. 27.

28. 29. 30.

31. 32. 33.

34. 35. 36.

Answer Options

37. 38. 39.

Select one or more options

[Clear All](#)

- The analyst of Chennai Super Kings wants to predict the number of sixes Mahi will hit in his first match of the IPL 2020 based on his performance in the IPL till now.
- The marketing team of upGrad wants to predict the number of new intakes in its Data Science Program for the next month.
- A company wants to predict its expenditure based on the cost of raw materials.
- The government of India wants to predict the rate of unemployment for the next quarter based on the data of this quarter.

62. 63.

64. 65. 66.

67. 68. 69.

70.

Coding

1. 2. 3.

4.

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

31 32 33

34 35 36

37 38 39

40 41 42

43 44 45

46 47 48

49 50 51

52 53 54

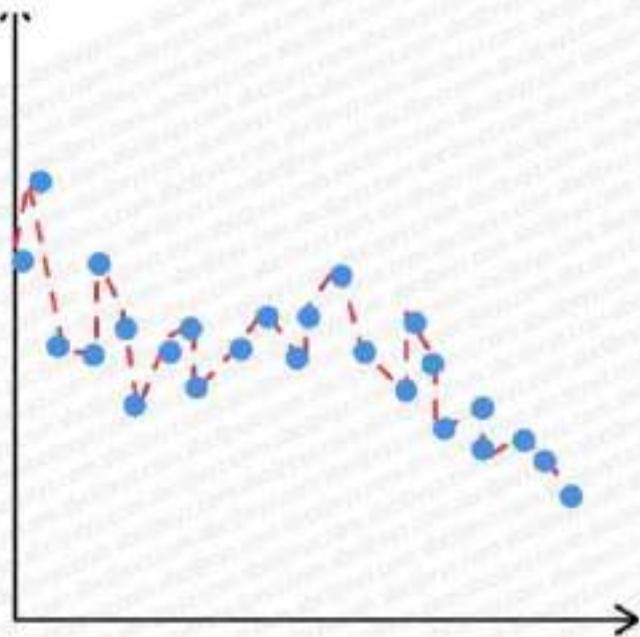
55 56 57

58 59 60

61 62 63

64 65 66

You are given an overfitted model as shown in the image given below. Which of the following options would reduce the complexity of the model, so that it aligns with the principle of bias-variance trade-off?



Answer Options

Select any one option

Clear All

- If the model is using a decision tree regressor, then the depth of the tree needs to be increased.
- Increasing the number of datapoints in the training set
- We need to have more variables in the data.

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

31 32 33

34 35 36

37 38 39

40 41 42

43 44 45

46 47 48

49 50 51

52 53 54

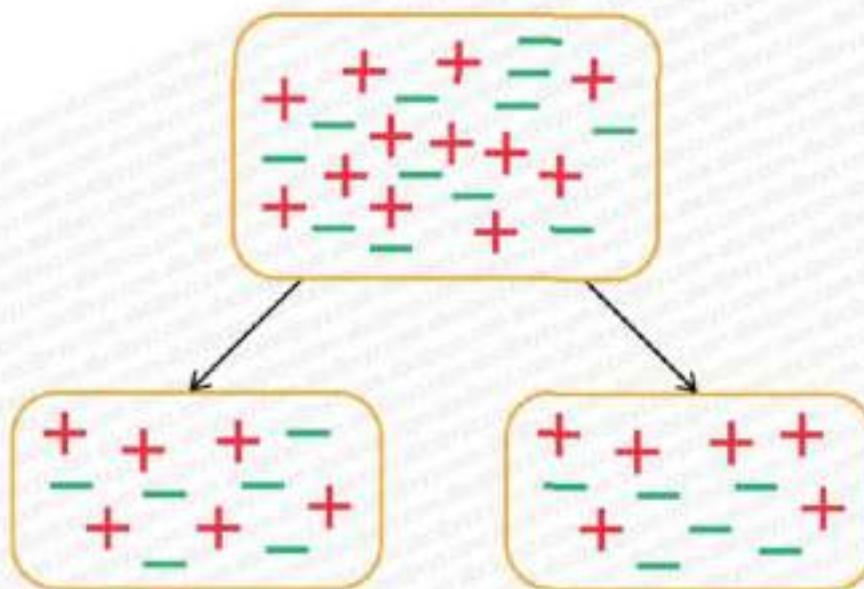
55 56 57

58 59 60

61 62 63

64 65 66

Refer to the following decision tree:



Consider the following statements.

Statement 1: The given tree is not a decision tree as both the leaf nodes are heterogeneous.**Statement 2:** The split is done incorrectly. The leaf nodes are as impure as the root node.

Answer Options

Select any one option

Clear A

 Statement 1 is correct and Statement 2 is incorrect Statement 1 is incorrect and Statement 2 is correct Both the statements are correct

Mode:

1	1	2
4	5	6
7	8	9

10	11	12
13	14	15
16	17	18

19	20	21
22	23	24
25	26	27

28	29	30
31	32	33
34	35	36

37	38	39
40	41	42
43	44	45

46	47	48
49	50	51
52	53	54

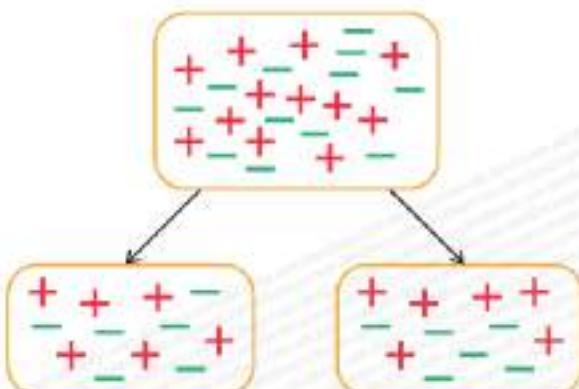
55	56	57
58	59	60
61	62	63

64	65	66
67	68	69
70		

Coding

1	2	3
4		

Refer to the following decision tree:



Consider the following statements:

Statement 1: The given tree is not a decision tree as both the leaf nodes are heterogeneous.
Statement 2: The split is done incorrectly. The leaf nodes are swapped as the root node.

Answer Options:

Select any one option

- Statement 1 is correct and Statement 2 is incorrect.
- Statement 1 is incorrect and Statement 2 is correct.
- Both the statements are correct.
- Both the statements are incorrect.

Q6. Q3 Multiple Correct Answer

MCQs

- 1 2 3
- 4 5 6
- 7 8 9
- 10 11 12
- 13 14 15
- 16 17 18
- 19 20 21
- 22 23 24
- 25 26 27
- 28 29 30
- 31 32 33
- 34 35 36
- 37 38 39
- 40 41 42
- 43 44 45
- 46 47 48
- 49 50 51
- 52 53 54
- 55 56 57
- 58 59 60
- 61 62 63
- 64 65 66
- 67 68 69
- 70

How is diversity achieved in the case of Random Forest? (More than one option may be correct.)

Answer Options

Select one or more options.

- Bootstrapped sampling is performed over the data to create multiple samples.
- We perform tree pruning for each tree built on the bootstrapped sample.
- A random subset of features are considered at each node of the tree built on the bootstrapped sample.
- A random subset of features are considered for building each tree.

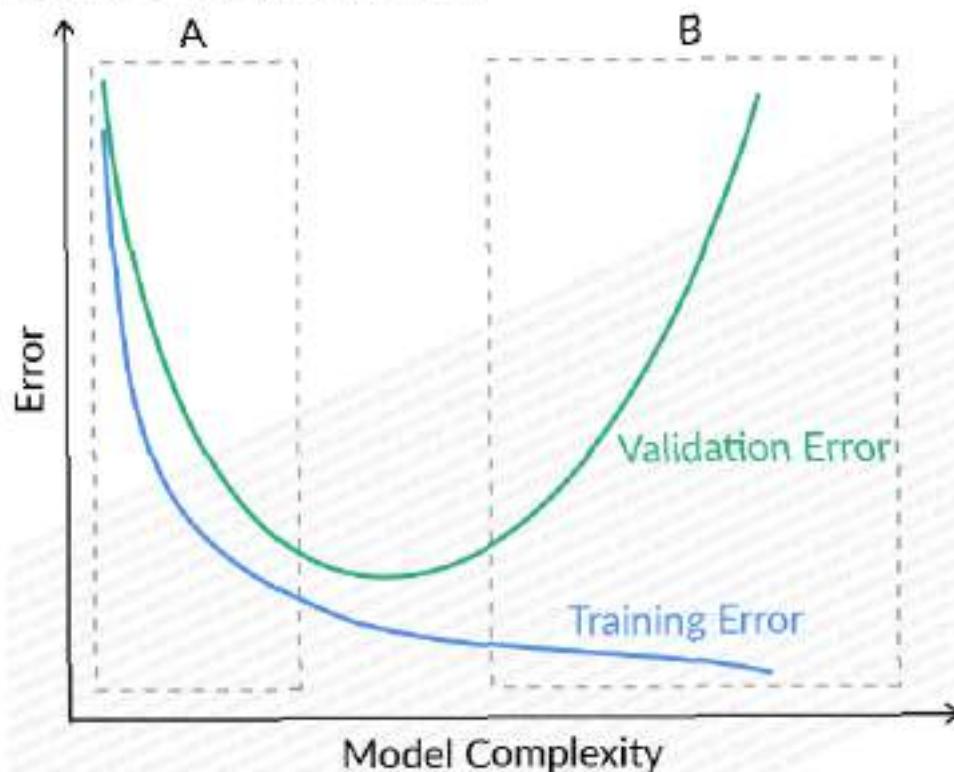
Coding

- 1 2 3
- 4

MCQs

1	2	3
4	5	6
7	8	9
10	11	12
13	14	15
16	17	18
19	20	21
22	23	24
25	26	27
28	29	30
31	32	33
34	35	36
37	38	
39	40	41
42	43	44
45	46	47
48	49	50
51	52	53
54	55	56
57	58	59
60	61	62
63	64	
65	66	67
68	69	70
71	72	73
74	75	76
77	78	79
80	81	82
83	84	85
86	87	88
89	90	91
92	93	94
95	96	97
98	99	100

Refer to site kaggle and choose the best option that represents line A and B.



Answer Options

Select any one option

 A- Underfitting, B- Good Model A- Good Model, B- Overfitting A- Underfitting, B- Overfitting A- Overfitting, B- Underfitting

175 min left

56. C3 Advanced ML

[Previous](#)[Next](#)

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

31 32 33

34 35 36

37 38 39

40 41 42

43 44 45

46 47 48

49 50 51

52 53 54

55

56

57

58

59

60

61

62

63

64

65

66

67

68

69

70

The following command is initialised in Python using the scikit-learn library for a decision tree model.

```
model=DecisionTreeClassifier(max_depth=4,min_samples_split=20,random_state=42)
```

Which of the following statements are true?

Answer Options

Select any one option.

[Clear A](#)

- The homogeneity metric used here is Gini.
- The random state is passed to make the output decision tree consistent.
- The minimum number of samples required to split an internal node is 20.
- All of the above.

Coding

1 2 3

4

175 min left

55. C3 Time Series Forecasting

[Previous](#)[Next](#)

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

31 32 33

34 35 36

37 38 39

40 41 42

43 44 45

46 47 48

49 50 51

52 53 54

55 56 57

58 59 60

61 62 63

Which is the compulsory component of any time series?

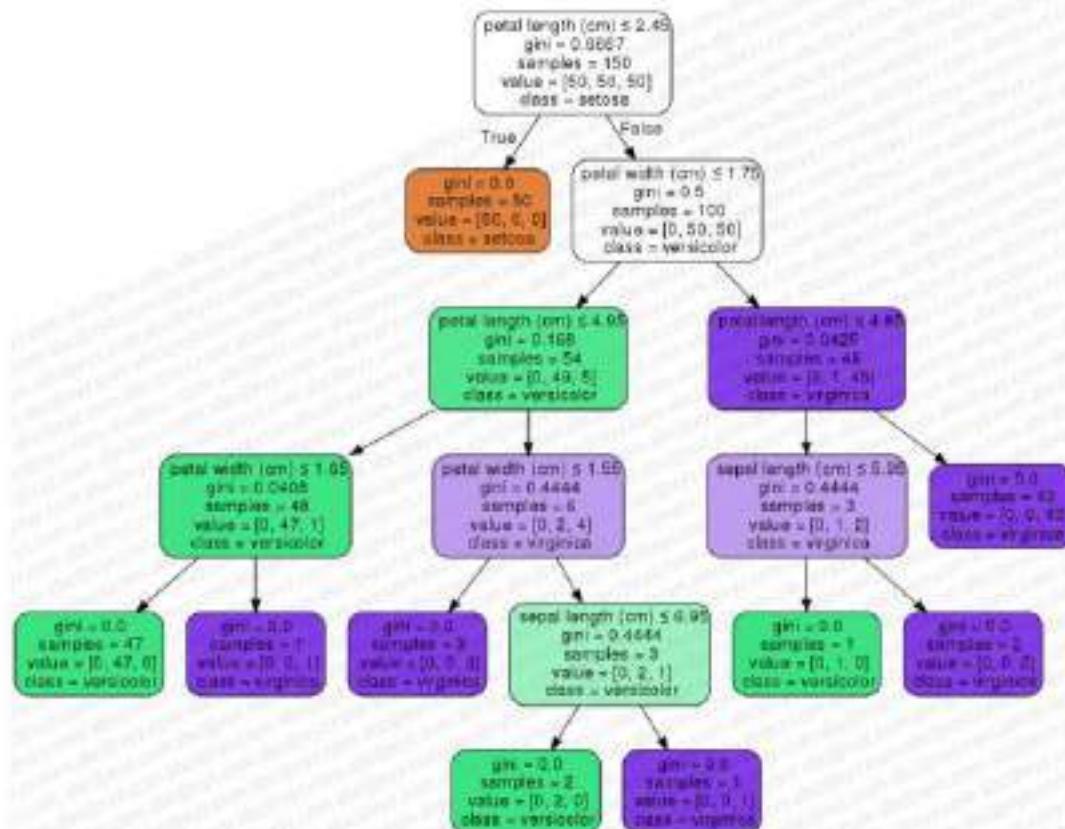
Answer Options

Select any one option

[Clear A](#) Level Trend Seasonality All of the above

1	2	3
4	5	6
7	8	9
10	11	12
13	14	15
16	17	18
19	20	21
22	23	24
25	26	27
28	29	30
31	32	33
34	35	36
37	38	39
40	41	42
43	44	45
46	47	48
49	50	51
52	53	54
55	56	57
58	59	60
61	62	63
64	65	66
67	68	69
70		

Refer to the decision tree given below and choose the statement that is correct as per this tree.



Answer Options

Select any one option

- The tree given above will show very good performance on the train data.
- The tree given above is an underfitting tree.
- If the petal length is more than 2.45, then it is equally likely that the flower is either setosa or virginica.

175 min left

53. C2-Decision Trees

[Previous](#)[Next](#)

25 26 27

28 29 30

31 32 33

34 35 36

37 38 39

40 41 42

43 44 45

46 47 48

49 50 51

52 53 54

55 56 57

58 59 60

61 62 63

64 65 66

67 68 69

70

Coding

1 2 3

Which of the following metrics measures how often a randomly chosen element would be incorrectly identified?

Answer Options

Select any one option

[Clear All](#) Entropy Information Gain Gini Index None of these

175 min left

52. C3 Time Series Forecasting

[Previous](#)[Next](#)

25 26 27

28 29 30

31 32 33

34 35 36

37 38 39

40 41 42

43 44 45

46 47 48

49 50 51

52 53 54

55 56 57

58 59 60

61 62 63

64 65 66

67 68 69

70

Coding

1 2 3

What does a MAPE value of 10% mean?

Answer Options

Select any one option

[Clear All](#)

- The absolute difference between the actual value and the forecasted value is 10%.
- The average absolute difference between the actual value and forecasted value is 10%.
- The forecasted value is 10% behind the actual value.
- The forecasted value is 10 times the actual value.

25 26 27

28 29 30

31 32 33

34 35 36

37 38 39

40 41 42

43 44 45

46 47 48

49 50 51

52 53 54

55 56 57

What will be the accuracy percentage of the given confusion matrix of the three-class classification?

True/ Predicted	Class A	Class B	Class C
Class A	13	0	5
Class B	0	4	8
Class C	1	1	9

Answer Options

Select any one option:

Clear Ans

 63% 36% 71% 45%

70

Coding

175 min left

50. C2 Logistic Regression

[Previous](#)[Next](#)

25 26 27

28 29 30

31 32 33

34 35 36

37 38 39

40 41 42

43 44 45

46 47 48

49 50 51

52 53 54

55 56 57

58 59 60

61 62 63

64 65 66

67 68 69

70

Coding

1 2 3

You have built a Logistic Regression model that is trying to predict whether a loan is approved or not based on a person's FICO score. Here are the model parameters: Intercept (B_0) = -9.346 and coefficient of FICO score = 0.0146. Given these parameters, can you calculate the probability of a loan getting approved for someone with a FICO score of 655?

Answer Options

Select any one option

[Clear All](#) 0.35 0.45 0.55 0.65

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

Which of the following strings will match with the regular expression '^01+0\$'?

- 1. 0
- 2. 00
- 3. 0111111110

Answer Options

Select any one option

Clear Ans

Only option 1

Only option 3

Both 1&2

Both 2&3



40 41 42

43 44 45

46 47 48

49 50 51

52 53 54

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

 0.137

31 32 33

 0.267

34 35 36

 0.527

37 38 39

 0.397

40 41 42

43 44 45

46 47 48

49 50 51

52 53 54

Random Forest		
Tree Number	Gini Impurity before split	Gini Impurity after split
1	0.39	0.31
2	0.46	0.28
3	0.40	0.21
4	0.42	0.32

Answer Options

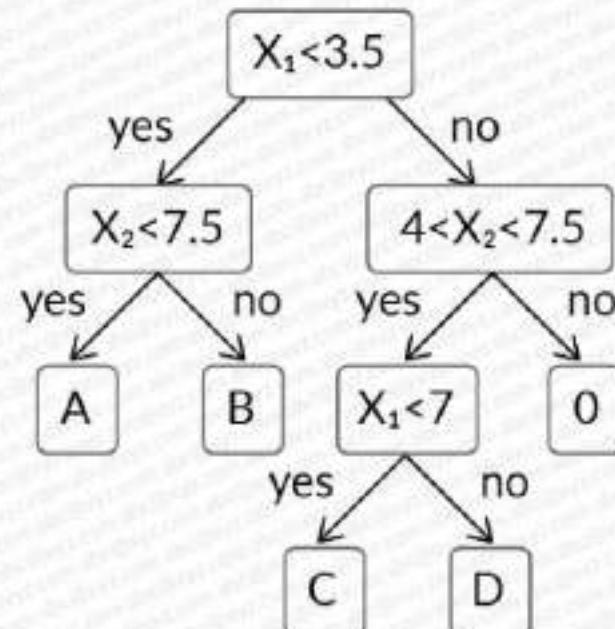
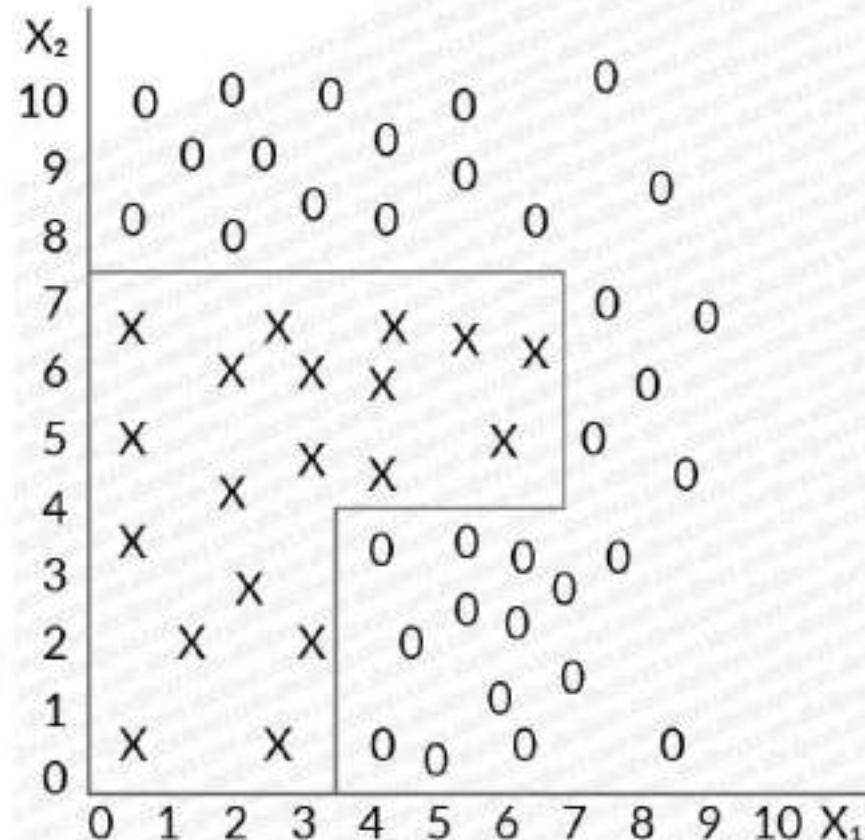
Select any one option

Clear Ans

MCQs

- 1 2 3
4 5 6
7 8 9
10 11 12
13 14 15
16 17 18
19 20 21
22 23 24
25 26 27
28 29 30
31 32 33
34 35 36
37 38 39
40 41 42
43 44 45
46 47 48
49 50 51
52 53 54
55 56 57
58 59 60
61 62 63
64 65 66

Refer to the image with the decision tree and the boundary diagram. Find what will be the outcome for the leaf nodes A, B, C, D.
Note: An outcome can be an "o" or an "x".



Answer Options

Select any one option

Clear All

 A: x, B: x, C: o, D: o A: o, B: o, C: x, D: x

MCQs:

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

31 32 33

34 35 36

37 38 39

40 41 42

43 44 45

46 47 48

49 50 51

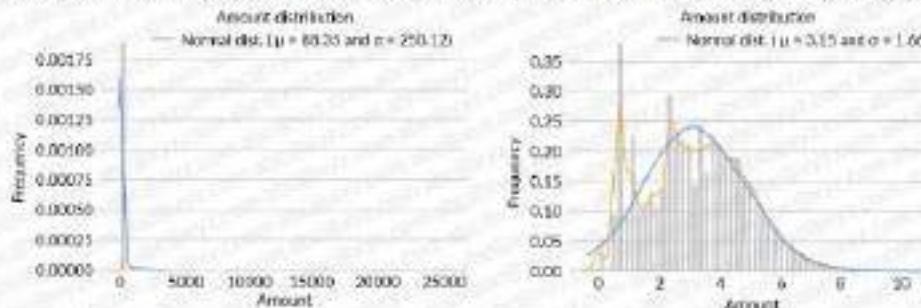
52 53 54

55 56 57

58 59 60

61 62 63

Refer to the plot below which consists of two plots A and B. Plot A represents the non-transformed column "Amount" while column B represents the transformed variable "Amount". Observe the plot B carefully and identify which kind of the transformation is applied to the variable "Amount".

**Answer Options**

Select any one option

Clear A

 Standardisation Log Transformation Min-Max Scaling Power Transformation

176 min left

40. C3- Advanced ML

[◀ Previous](#)[Next ▶](#)

MCQs:

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

31 32 33

34 35 36

37 38 39

40 41 42

43 44 45

46 47 48

49 50 51

52 53 54

55 56 57

58 59 60

61 62 63

Feature engineering is an important step in any model building exercise. It is the process of creating new features from a given data set using the domain knowledge to leverage the predictive power of a machine learning model. Which of the following statements are correct?

Statement 1: Feature engineering techniques are applied before train-test split.

Statement 2: There is no difference between standardization and normalization.

Statement 3: Mean encoding is a feature engineering technique for handling categorical features.

Answer Options

Select any one option

Clear A

 Only 1 and 2 Only 2 and 3 Only 1 Only 3

176 min left

39. C3 Time Series Forecasting

[Previous](#)[Next](#)

MCQs:

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

Answer Options

22 23 24

Select any one option.

[Clear A](#)

25 26 27

 6

28 29 30

 5

31 32 33

 3

34 35 36

 None of these

37 38 39



40 41 42

43 44 45

46 47 48

49 50 51

52 53 54

55 56 57

58 59 60

61 62 63

176 min left

38. C2 linear Regression

[Previous](#)[Next](#)

MCQs:

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

What does standardised scaling do?

22 23 24

Answer Options

25 26 27

Select any one option.

[Clear A](#)

- Bring all data points in the range 0 to 1
- Bring all data points in the range -1 to 1
- Bring all the data points in a normal distribution with mean 0 and standard deviation 1
- Bring all the data points in a normal distribution with mean 1 and standard deviation 0

31 32 33

34 35 36



37 38 39

40 41 42

43 44 45

46 47 48

49 50 51

52 53 54

55 56 57

58 59 60

61 62 63

176 min left

37. C2 Clustering

[Previous](#)[Next](#)

MCQs:

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

Which of the following is not true for Hopkins statistic?

22 23 24

Answer Options

25 26 27

Select any one option.

[Clear A](#) Hopkins statistic decides if the data is suitable for clustering or not

28 29 30

 Hopkins statistic lie between -1 and 1

31 32 33

34 35 36

 If the Hopkins statistic comes out to be 0, then the data is uniformly distributed

37 38 39

40 41 42

 If the Hopkins statistic comes out to be 1, then the data highly suitable for clustering

43 44 45

46 47 48

49 50 51

52 53 54

55 56 57

58 59 60

61 62 63

177 min left

36. C3 Advanced ML

< Previous

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

31 32 33

34 35 36

<<

37 38 39

40 41 42

43 44 45

46 47 48

49 50 51

52 53 54

56 56 57

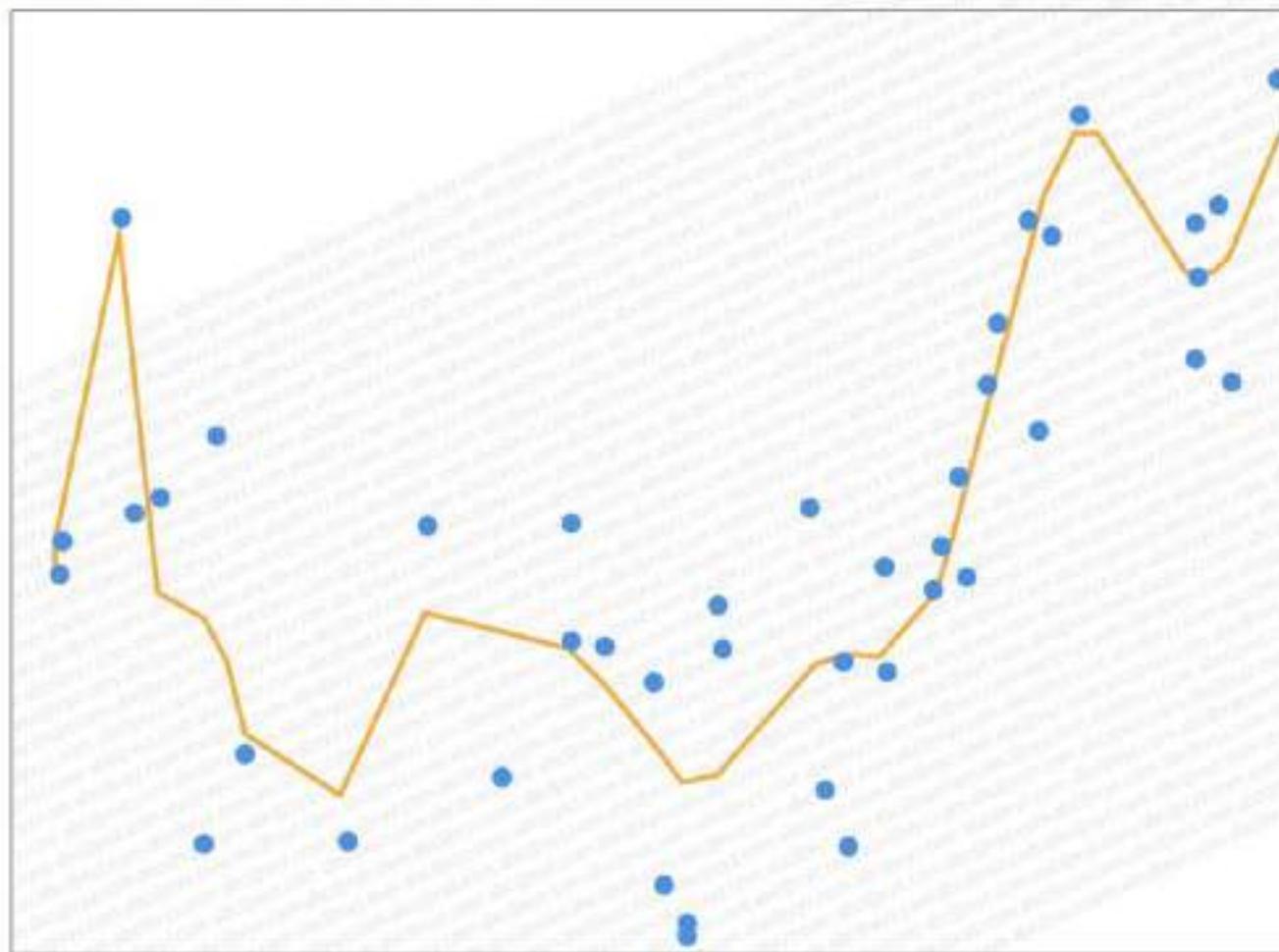
58 59 60

61 62 63

64 65 66

67 68 69

70



Answer Options

Select any one option

 He will get an underfitting model as the increase in the polynomial degree will not help reduce the model bias. He will get an overfitting model as the current model also overfits the train data.

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30



31 32 33

34 35 36



37 38 39

40 41 42

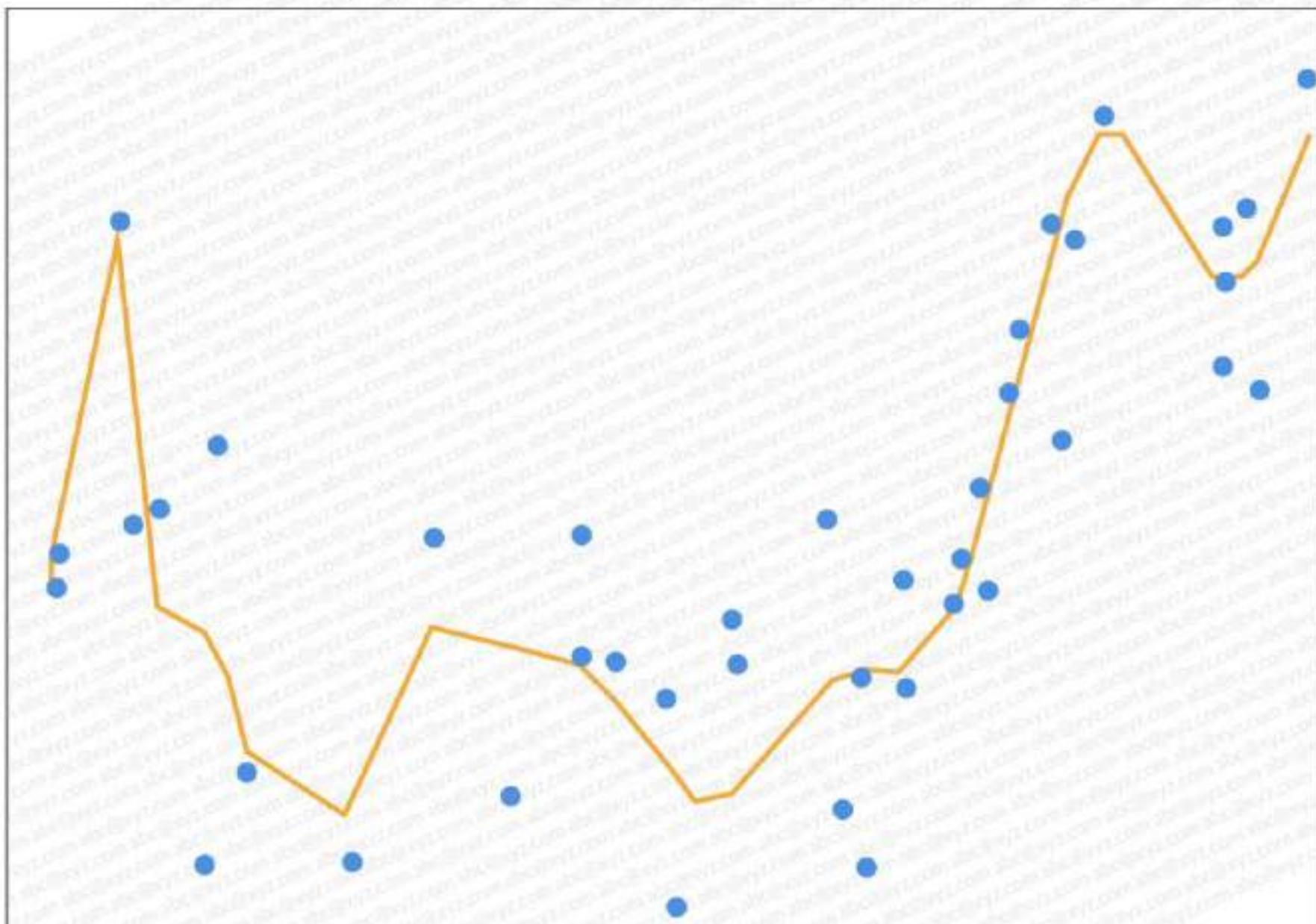
43 44 45

46 47 48

49 50 51

52 53 54

Joy is a budding data scientist and he is currently working on a polynomial regression model. The model that he has created has a polynomial degree of 3 and it is depicted in the image given below. To increase the accuracy of the model and reduce error on train data, he decides to increase the degree of polynomial to 5. Choose the best option, which will represent his model with a polynomial degree of 5.



177 min left

35. C2 Logistic Regression

[Previous](#)[Next](#)

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

Consider the following confusion matrix.

Total=500	Actual Positive	Actual Negative
Predicted Positive	196	20
Predicted Negative	28	296

Which among the following is the lowest for the given confusion matrix?

Answer Options

Select any one option

[Clear Ans](#) Accuracy Precision Sensitivity Specificity

31 32 33

34 35 36

37 38 39

40 41 42

43 44 45

46 47 48

49 50 51

52 53 54

177 min left

34. C3 Advanced ML

[Previous](#)[Next](#)

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

Which of the following statements is NOT true for the decision tree regression?

Answer Options

Select any one option

[Clear Ans](#)

- Leaves in decision tree regression contain average values as the prediction.
- Impurity measure for a given node is measured by the weighted mean square error.
- In decision tree regression, a lower value of mean square error means that the data values are dispersed widely around mean.
- Weighted mean square error is nothing but the variance of the observations.



35 36

37 38 39

40 41 42

43 44 45

46 47 48

49 50 51

52 53 54

177 min left

33. C2 Logistic Regression

[Previous](#)[Next](#)

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

Consider the following univariate logistic model:

$$Y = \beta_0 + \beta_1 X_1$$

Which of the following statements is NOT true?

Answer Options

Select any one option

[Clear All](#)

- The maximum likelihood estimation determines the best combination of β_0 and β_1 .
- If β_1 is increased by 1 unit, Y increases by 1 unit.
- β_0 is the y-intercept
- If β_1 is increased by 1 unit, log-odds increases by 1 unit.



33

34 35 36

37 38 39

40 41 42

43 44 45

46 47 48

49 50 51

52 53 54

177 min left

32. C3- Advanced ML

[◀ Previous](#)[Next ▶](#)

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

Answer Options

Select any one option

[Clear Ans](#) 1:5 2:3 1:7 None of these

31 32 33

34 35 36

37 38 39

40 41 42

43 44 45

46 47 48

49 50 51

52 53 54

177 min left

31. C3 Time Series Forecasting

[Previous](#)[Next](#)

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

31 32 33

34 35 36

37 38 39

40 41 42

43 44 45

46 47 48

49 50 51

52 53 54

You are an analyst at a retail company in India. Owing to the COVID-19 pandemic, there is a huge demand for sanitisers, and the rate of infection is following an upward trajectory (a higher number of infections daily). You have access to past 30 days data for the demand of sanitisers. Which of the following models would work best for forecasting the demand for sanitisers for the next 5 days with highest accuracy?

Answer Options

Select any one option

[Clear All](#)

- Seasonal Auto-Regressive Integrated Moving Average Model
- Simple Average Forecast Method
- Simple Exponential Smoothing Method
- Holt-Winters' Method

MCQs

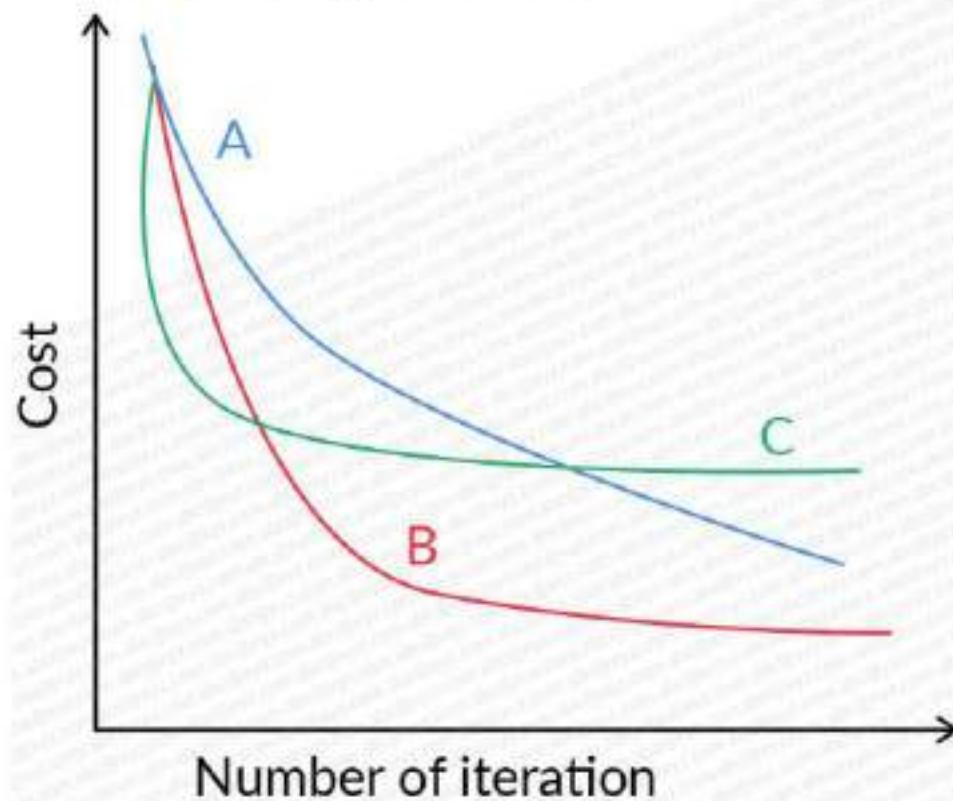
- 1 2 3
- 4 5 6
- 7 8 9
- 10 11 12
- 13 14 15
- 16 17 18
- 19 20 21
- 22 23 24
- 25 26 27

30

- 28 29 30
- 31 32 33
- 34 35 36
- 37 38 39
- 40 41 42
- 43 44 45
- 46 47 48
- 49 50 51



Observe the following cost function graph with different learning rates.



Which of the following statements are true about the learning rate? (More than one option may be correct.)

Answer Options

Select one or more options

Clear

The learning rate of curve C is highest among all curves

The learning rate for curve B is lower than A

The learning rate for curve B is higher than A

Coding

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

25 26 27

28 29 30

Answer Options

Select any one option

Clear Ans

 0.1 0.2

Observations	Trees Predictions				Actual Label	Predicted Label
	A	B	C	D		
3	-	-	-	0	-	
10	0	0	0	-	0	
5	-	-	0	-	0	
2	-	-	-	0	-	
8	0	0	0	-	0	
1	0	-	0	0	-	
6	-	-	-	-	-	
9	0	0	0	0	0	
4	-	-	-	0	0	
7	0	0	0	-	0	

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

<<

25 26 27

28 29 30

31 32 33

34 35 36

37 38 39

Refer to the table below and find the OOB Error for a Random Forest model that consists of 4 trees namely A, B, C, D and 10 observations. Use the table below and fill the column "Predicted Label" and then make use of it to calculate the OOB error.

Class Labels:

"O": Positive Class

"-": Negative Class

Observations	Trees Predictions				Actual Label	Predicted Label
	A	B	C	D		
3	-	-	-	O	-	-
10	O	O	O	-	O	O
5	-	-	O	-	O	O
2	-	-	-	O	-	-
8	O	O	O	-	O	O
1	O	-	O	O	-	-
6	-	-	-	-	-	-
9	O	O	O	O	O	O
4	-	-	-	O	O	O
7	O	O	O	-	O	O

Answer Options



28. C2 Logistic Regression

[◀ Previous](#)[Next ▶](#)

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

<<

25 26 27

28 29 30

31 32 33

34 35 36

37 38 39

30 31 32

Given an imbalanced dataset, the ratio of positive to negative class is 1:10000. You run a logistic regression model and find out that the model has a high value of precision and a low value of recall. Which of the following statements is true?

Answer Options

Select any one option

[Clear Ans](#)

- The class is handled well by the data
- The model is not able to detect the class, but when it does it is highly trustable
- The model is able to detect the class but it includes data points from the other class as well
- The class is handled poorly by the data

177 min left

27. C2 Clustering

[Previous](#)[Next](#)

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

<<

25 26 27

27

28 29 30

31 32 33

34 35 36

37 38 39

35 36 37

Initialising the following command in Python will result in the following: `model_clus = KMeans(n_clusters = 6, max_iter=50)`

Answer Options

Select any one option

[Clear Ans](#)

- Run maximum 6 iterations
- Run maximum 40 iterations
- Create 6 final clusters
- Create 50 final clusters

177 min left

26. C2 Logistic Regression

[Previous](#)[Next](#)

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

<<

25 26 27

28 29 30

31 32 33

34 35 36

37 38 39

30 31 32

Which of the following is correct for a logistic regression model?

Answer Options

Select any one option

[Clear Ans](#)

- The independent variables should not be multicollinear.
- The dependent variable should follow Normal Distribution.
- The log odds in a logistic regression model lies between 0 and 1.
- F1-score is always the best metric for evaluating a logistic regression model.

177 min left

25. C2 linear Regression

[Previous](#)[Next](#)

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

<<

25 26 27

28 29 30

31 32 33

34 35 36

37 38 39

Which of the following is true regarding the error terms in linear regression?

Answer Options

Select any one option

[Clear Ans](#)

- The sum of residuals should be zero
- The sum of residuals should be lesser than zero
- The sum of residuals should be greater than zero
- There is no such restriction on what the sum of residuals should be

24. C2 Logistic Regression

[Previous](#)[Next](#)

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

Which of the following is true for weight of evidence (WoE) analysis?

[Clear Ans](#)

Answer Options

Select any one option

- It helps in finding the different predictive patterns for the different segments that might be present in the data.
- WoE helps in treating missing values for both continuous and categorical variables.
- WoE values should follow an increasing or decreasing trend across bins.
- All of the above

35 36

37 38 39

35 36

37 38 39

35 36

37 38 39

177 min left

24. C2 Logistic Regression

[Previous](#)[Next](#)

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

Which of the following is true for weight of evidence (WoE) analysis?

Answer Options

Select any one option

 Proctoring violations detected[Clean Answer](#)

It helps in finding the different predictive patterns for the different segments that might be present in the data. You must not leave the frame of the camera for the duration of the test.

All violations will be recorded and visible in your report.

WoE helps in treating missing values for both continuous and categorical variables.

[Continue test](#)

WoE values should follow an increasing or decreasing trend across bins.

All of the above

23. C2 Multiple correct answer

[Previous](#)[Next](#)

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

<<

25 26 27

28 29 30

31 32 33

34 35 36

37 38 39

In a simple linear regression model when you fit a straight line through the data you'll get the two parameters of the straight line, i.e. the intercept β_0 and the slope β_1 . Which of the following is true for β_0 and β_1 ? (More than one option may be correct.)

Answer Options

Select one or more options

[Clear Ans](#) The null hypothesis for a simple linear regression model is $H_0: \beta_1 = 0$ If the p-value turns out to be greater than 0.05 for β_1 , it means β_1 is significant If β_1 turns out to be insignificant, that means there is no relationship between the dependent and independent variable If the p-value turns out to be less than 0.05 for β_0 , it means that β_0 is non-zero

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

<<

25 26 27

28 29 30

31 32 33

34 35 36

37 38 39

Which of the following describes an advantage of the random forest (RF) algorithm over the decision tree (DT) algorithm?

Answer Options

Select any one option

Clear Ans

- An RF model has better interpretability as compared to a DT
- An optimally fit RF model has less variance than an optimally fit DT
- An optimally fit RF model has more variance than an optimally fit DT
- Building an RF model is computationally less expensive than building a DT

178 min left

21. C3 Advanced ML

[Previous](#)[Next](#)

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

<<

25 26 27

28 29 30

31 32 33

34 35 36

37 38 39

40 41 42

Which of the following methods is NOT true for the truncation of a decision tree?

Answer Options

Select any one option

[Clear Ans](#)

- Limit the depth of a tree
- Set a minimum threshold on the number of samples that appear in a leaf.
- Merging of two non-leaf nodes.
- Truncation is also known as pre-pruning.



20. C2 linear Regression

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

<<

25 26 27

28 29 30

31 32 33

34 35 36

37 38 39

40 41 42

In linear regression, the metric F-statistic is used to determine

Clear Ans

Answer Options

Select any one option

 the significance of the individual beta coefficients the variance explanation strength of the model the significance of the overall model fit Both A & C

19. C3 Multiple Correct Answer

[◀ Previous](#)[Next ▶](#)

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24



Answer Options

25 26 27

[Clear Ans](#)

28 29 30

Select one or more options

31 32 33

 A: 15.5

34 35 36

 B: -1.25

37 38 39

 B: 1.75

18. C2 linear Regression

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

<<

25 26 27

28 29 30

31 32 33

34 35 36

37 38 39

40 41 42

Consider the following two assumptions for a simple regression model. (Assume X and y to be independent and dependent variables respectively).

Statement 1: There is a linear relationship between X and y.

Statement 2: X and y are normally distributed.

Answer Options

Select any one option

Clear Ans

 Statement 1 is correct and Statement 2 is incorrect Statement 1 is incorrect and Statement 2 is correct Both the statements are correct Both the statements are incorrect

17. C2 Clustering

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

Which of the following statements is NOT true?

Answer Options

Select any one option

Clear Ans

- Each time the clusters are made during the K-means algorithm, the centroid is updated.
- The cluster centres that are computed in the K-means algorithm are given by centroid value of the cluster points.
- Standardization of the data is not important before applying Euclidean distance as a measure of similarity/dissimilarity
- The centroid of a column with data points 25, 32, 34 and 23 is 28.5

178 min left

16. C2 Logistic Regression

[◀ Previous](#)[Next ▶](#)

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

<<

25 26 27

28 29 30

31 32 33

34 35 36

37 38 39

40 41 42

If you use a random number generator to predict the output 0 or 1 for a binary classification problem, what will be the area under the curve of the ROC curve?

Answer Options

Select any one option

[Clear Ans](#) 0 0.5 1 100

178 min left

16. C2 Logistic Regression

[← Previous](#)[Next →](#)

MCQs

If you use a random number generator to predict the output 0 or 1 for a binary classification problem, what will be the area under the curve of the ROC curve?

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

Answer Options

16 17 18

Select any four option

19 20 21

0

22 23 24 25

0

26 27

0.5

28 29 30

0.5

31 32 33

1

34 35 36

100

37 38 39



Proctoring violations detected

No face detected

You must not leave the frame of the camera for the duration of the test

All violations will be recorded and visible in your report

[Continue test](#)

15. C2 Business Problem Solving

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

<<

 Neural Network

25 26 27

 Logistic regression

28 29 30

 Decision tree

31 32 33

 All of the above

34 35 36

37 38 39

The coronavirus disease (COVID-19), was declared a pandemic by the World Health Organization (WHO) in February 2020. Currently, there are no vaccines or treatments that have been officially approved by WHO after clinical trials. India has not seen the peak of infection yet and the number of infections is touching a new height daily. The business unit of an Indian health and hygiene company approaches you to know "Why the sales of masks is decreasing despite the number of corona infections increasing daily".

Answer the following questions:

Suppose you mapped the above problem statement with a classification problem, either a customer will buy a mask or not. You will build _____ model as your initial solution.

Answer Options

Select any one option

Clear Ans

 Neural Network Logistic regression Decision tree All of the above

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24



25 26 27

28 29 30

31 32 33

34 35 36

37 38 39

40 41 42

We wanted to tune hyperparameters for a Random Forest model using Grid Search technique with the CV of 5 folds. Refer to the list of hyperparameters that are required to be tuned.

```
{'n_estimators': [10, 25], 'max_features': [5,10],  
 'max_depth': [10, 50, None], 'bootstrap': [True, False]  
}
```

Assume that each set of hyperparameters takes 1 minutes to run, find the total time required to complete the tuning process.

Answer Options

Select any one option

Clear Ans

 60 minutes 45 minutes 90 minutes 120 minutes

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

What would happen if you choose a very high value of the hyperparameter, lambda?

$$\min_{a,b} \left[\sum_{i=1}^n (y_i - ax_i - b)^2 + \lambda(a^2 + b^2) \right]$$

Answer Options

Select any one option

Clear Ans

 The model would become simpler, yet it would show robust performance on test data. The model would become too simple. It would become an underfitted model. The model would become too complex. It would become an overfitted model.

12. C2 Business Problem Solving

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24



Answer Options

Select any one option

Clear Ans

 Statement 1 is correct and Statement 2 is wrong Statement 2 is correct and Statement 1 is wrong Both the statements are correct

178 min left

11. C2 linear Regression

[Previous](#)[Next](#)

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

<<

25 26 27

28 29 30

31 32 33

34 35 36

37 38 39

How is regression different from classification?

Answer Options

Select any one option

[Clear Ans](#)

- One is supervised while the other is unsupervised
- One is iterative while the other is closed form
- In regression, the response variable is numeric while it is categorical in classification
- None of the above

10. C2 Business Problem Solving

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24



Answer Options

25 26 27

Select any one option

Clear Ans

28 29 30

 3>1>5>2>4>7>6>8

31 32 33

 3>2>1>5>4>7>6>8

34 35 36

 4>3>1>2>5>7>6>8

37 38 39

 3>2>1>5>4>7>8>6

40 41 42



9. C2 Multiple correct answer

[◀ Previous](#)[Next ▶](#)

MCQs

1 2 3

4 5 6

7 8 9

10 11 12

13 14 15

16 17 18

19 20 21

22 23 24

<<

25 26 27

28 29 30

31 32 33

34 35 36

37 38 39

ROC curve shows the tradeoff between the True Positive Rate (TPR) and the False Positive Rate (FPR). TPR and FPR are sensitivity and (1 - specificity) respectively. The following function is written in Python using metrics package from the scikit-learn library for a ROC curve function.

```
def draw_roc(actual, probs):
    fpr tpr,thresholds=metrics.roc_curve(actual,probs,drop_intermediate=False)
    auc_score = metrics.roc_auc_score(actual, probs)
    return None
```

Which of the following statements are true? (More than one option may be correct.)

Answer Options

Select one or more options:

[Clear Ans](#) The area under the ROC curve can be more than 1. The arguments passed in the above function are actual values of the target variable and the predicted values (i.e., 0 or 1) Larger the area under the curve, the better will be the model

Question 1.

Given a list of integers and another integer 'k', find the kth largest integer in the list. If there are less than k' distinct elements in the list, then you need to output - 1. (Refer to the sample inputs and outputs for details.).

Note: k=1 means you have to find the largest element

Input Format:

Line 1 contains a list of integers

Line 2 contains a positive integer $k > 0$

Output Format:

An integer, kth largest integer

Examples:

Sample Input 1:

[2, 3, 1, 5, 6, 2, 1]

4

Sample Output 1:

2

Sample Input 2:

[2, 3, 1, 5, 6, 2, 1]

6

Sample Output 2:

-1

Question 2.

Silhouette metric for any ith point is given by $S(i) = (b(i) - a(i)) / \max(a(i), b(i))$

Which of the following is not true about the Silhouette metric?

- a. $b(i)$ is the average distance from the nearest neighbor cluster_____
- b. $a(i)$ is the average distance from own cluster
- c. If $S(i) = 1$ the the data point is similar to its own cluster
- d. Silhouette metric ranges from 0 to +1.....

Question 3.

A client approached you with a problem statement. You decided to build a multiple linear regression model on the dataset provided. The dataset consisted of 40 features. Obviously all features will not be significant. Selecting the relevant features manually will be a tougher task. You can use RFE to select relevant features. RFE is an automated feature selection technique. Initially, you assumed 25 fe can explain your whole data.

Which of the following commands correctly calls the RFE technique in Python? (Here is the 'lm' is the fitted instance of multiple linear regression model)

- a. from statsmodel.feature_selection import RFE
rfe=RFE(lm, 25)
rfe = rfe.fit(X_train, y_train)
- b. from sklearn.feature_selection import RFE
rfe=RFE(lm, 25)
rfe = rfe.predict(X_train, y_train)
- c. from sklearn.feature_selection import RFE..... 
rfe=RFE(lm, 25)
rfe = rfe.fit(X_train, y_train)
- d. from RFE import feature_selection
rfe=RFE(lm, 25)
rfe = rfe.predict(X_train, y_train)

Question 4.

Some of the independent variables (predictors) might be interrelated due to which the presence of a particular independent variable in the model is redundant. This phenomenon is called Multicollinearity. Suppose that you are building a multiple linear regression model for a given problem statement. Which of the following statements is TRUE w.r.t. Multicollinearity?

- a. Multicollinearity is a problem when your only goal is to predict the independent variable from a set of dependent variables
- b. Multicollinearity is a problem when your goal is to infer the effect on the dependent variable due to independent variables

- c. Multicollinearity is not a problem if a variable is not collinear with your variable of interest
- d. Multicollinearity is not a problem if there are multiple dummy (binary) variables that represent a categorical variable with three or more categories

Question 5.

Which of the following assumptions do we make while building a simple linear regression model? (assume X and y to be independent and dependent variables respectively).

- A. There is a linear relationship between X and y
- B. X and y are normally distributed
- C. Error terms are independent of each other
- D. Error terms have constant variance.
 - a. A, B, C and D
 - b. A, C and D.....
 - c. A, B and C
 - d. B, C and D

Question 6.

The output of an Logistic model is

- a. 0 or 1
- b. Any value between 0 and 1.....
- c. 0.5
- d. Depends on business problems

Question 7.

What will be the output of :

Pattern = '\w+ed'

String = "He played and won the match when he was injured"
re.search(Pattern, String)

- a. Played, injured

- b. Injured
- c. Played.....
- d. It will throw an error

Question 8.

Consider the following confusion matrix.

<col width="100"> <col width="99"> <col width="101">

Total = 500	Actual Positive	Actual Negative
Predictive Positive	196	20
Predicted Negative	28	256

Which among the following is the highest for the given confusion matrix?

- a. Sensitivity
- b. Specificity
- c. Precision
- d. Accuracy

Question 9.

Consider the following two statements

Statement 1: Suppose the value of Precision and Recall for a model are 0.65 and 0.75 respectively. Then the value of F1-score will be -0.696

Statement 2: Mean squared error is a metric that can be used to evaluate a logistic regression model

- a. Statement 1 is wrong and statement 2 is correct
- b. Statement 1 is correct and statement 2 is wrong
- c. Both the statements are correct
- d. None of the statements are correct

Question 10.

Which of the following is NOT true?

- a. In the case of a fair coin, the odds of getting heads is 1
- b. The error values of linear and logistic regression have to be normally distributed
- c. Specificity decreases with an increase in sensitivity
- d. As TPR increases, FPR also increases

Question 11.

An analyst at a multinational e-commerce firm is trying to visualise the number of units of different items that were sold in the past financial year. They want to display these items in a single bar chart. However, since a few items (being daily-use items) had sales quantity in the millions and a few other premium items had sales quantity in single digits, the analyst was facing difficulty accommodating all of them in a single graph.

Which of the following suggestions would be the most helpful for the analyst to achieve the desired outcome?

- a. They should use a stacked bar chart
- b. They should use two separate bar charts
- c. **They should use a dual-axis bar chart.....**
- d. They should use a horizontal bar chart.

Question 12.

Mr. X has created a Tableau workbook and wants to save only the visualisations and then share the workbook with Mr. Y for discussions. Which of the following file formats should Mr X use in order to save his workbook?

- a. .tbl
- b. **.twb.....**
- c. .tblx
- d. .twbx

Question 13.

An analyst wants to visually display the country wise breakup of COVID-19 cases across the world. They want to depict (in terms of percentage) the countries with the highest proportion of global cases.

Which of the following graphs would be best suited for this scenario?

- | | |
|---|---|
| a. Stacked Bar Plot | a. Bar Plot |
| b. TreeMap | b. Line Chart |
| c. Line Plot | c. Heat Map |
| d. Pie Chart.  | d. Pie Chart.  |
- OR

Question 14.

The analyst from the previous question wants to visually depict the percentage of male and female COVID-19 cases for each country as well.

Which of the following graphs will be best suited for this scenario?

- a. Stacked Bar Plot..... 
- b. TreeMap
- c. Line Plot
- d. Gantt Chart

Question 15.

The coronavirus disease (COVID-19), was declared a pandemic by the World Health Organization (WHO) in February 2020. Currently, there are no vaccines or treatments that have been officially approved by WHO after clinical trials, India has not seen the peak of infection yet and the number of infections is touching a new height daily. The business unit of an Indian health and hygiene company approaches you to know "Why the sales of masku is decreasing despite the number of corona infections increasing daily".

Answer the following questions

Suppose you mapped the above problem statement with a classification problem, either a customer will buy a mask or not. You will build _____ model as your initial solution.

- a. Neural Network
- b. Logistic Regression
- c. Decision Tree
- d. All of the above

Question 16.

After performing PCA if the first principal component explains about 41% of the variance of the entire dataset. Which of the following values could represent the variance explained by the second principal component? (More than one option may be correct)

- a. 56%
- b. 33%.....
- c. 41%
- d. 64%

Question 17.

Regarding bias and variance, select the correct statement/s (Here high and low are relative to the ideal model) (MCQ)

- a. Models which overfit have a high bias
- b. Models which overfit have a low bias.....
- c. Models which underfit have a high variance
- d. Models which underfit have a low variance.....

Question 18.

Which of the following properties are characteristics of a decision tree? (MCQ)

- a. High bias
- b. High variance.....
- c. Lack of smoothness of prediction surfaces.....
- d. None of the above

Question 19.

Choose the correct option based on the following statements about Tableau (More than one option may be correct)

Statement 1: In a box plot, the whisker denote the 25th and 75th percentiles of the data

Statement 2: Pivoting is the process of converting rows to columns, and vice versa, in a data set

Statement 3: A treemap can accommodate two measures only: One to control the size and the other to control colour

- a. Statement 1 is True and statement 2 is False
- b. Statement 1 is False and statement 2 is True
- c. Statement 2 is False and statement 3 is True
- d. Statement 2 is True and statement 3 is True..... 

Question 20.

How to Attempt?

Write a Python Function to check whether a number is perfect or not. A perfect number is a positive integer that is equal to the sum of its proper positive divisors, that is, the sum of its positive divisor excluding the number itself

Example: The first perfect number is 6, because 1, 2 and 3 are its proper positive divisors, and $1+2+3 = 6$. The next perfect number is $28 = 1+2+4+7+14$, and so on.

Input : A positive integer

Output : 1 if the number is a perfect number else output 0.

Input 1:

6

Output 1:

1

Input 2:

10

Output 2:

0

Solution 20.

```
sum = 0
for x in range(1, input1):
    if input1 % x == 0
        sum += x
return sum == input1
```

Question 21.

Which of the following is the hyperparameter used in regularised regression?

$$\text{RSS} = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^p \beta_j^2$$

- a. Lambda..... 
- b. Alpha
- c. None of the above
- d. All of the above

Question 22.

In which of the following algorithms, do we not use a decision boundary to classify the data points?

- a. Logistic Regression
- b. Decision Tree
- c. Both a and b
- d. None of the above

Question 23.

A piece of strong evidence for the generalizability of ensemble models such as random forests over standalone models like logistic regression is that they

- a. Show good cross validation performance
- b. Use lower computational power
- c. Exhibit low model variance
- d. All of the above

Question 24.

The primary issue with models having high variance is

- a. Model produced is too sensitive to the input data, it will also learn what shouldn't be learnt from the data given
- b. The model becomes too complex to deal with

- c. Takes a lot of time and computational effort to train such model
- d. All of the above

Question 25.

Cross-validation results indicate a good choice of the model when it shows

- a. High average accuracy with high variance
- b. Low average accuracy with low variance
- c. High average accuracy with low variance
- d. Low average accuracy with high variance

Question 26.

Suppose we have an overfitted model which of the following is NOT a valid method to reduce the overfitting?

- a. Decrease the model complexity
- b. Improve the optimisation algorithm being used for error minimization
- c. Increase the amount of training data
- d. Regularisation

Question 27.

Take a look at the following three problem statements.

Problem statement 1: Let's say that you are building a telecom churn prediction model with the

Business objective that your company wants to implement an aggressive customer retention campaign to retain the high churn risk customers. This is because a competitor has launched extremely low cost mobile plans and you want to avoid churn as much as possible by incentivising the customers. Assume that budget is not a constraint.

Problem statement 2: Let's say you are building a cancer detection model with the objective that both the patient who had cancer and the patient who has not cancer can be detected correctly. It can have serious implications if you predict either of the class wrong, i.e. if wrongly detected as "not cancer", the patient will die of cancer, if wrongly detected as "cancer", the patient will die of chemotherapy.

Problem statement 3: You have build an image classification model where 60% of images belong to one class and the rest 40% images belong to another class. You have to predict the class of a new image

Which is the correctly matched model evaluation metric for the above classification models?

- a. Problem statement 1: Specificity
- b. Problem statement 2: Sensitivity
- c. Problem statement 3: Specificity
- d. Problem statement 4: Accuracy

Question 28.

Identify the reasons that justify the usage of PCA before regression and choose the correct option

- P. To find the information missing from the input data
- Q. For model explainability, we prefer features that are orthogonal
- R. To reduce multicollinearity
- S. To make computation faster by reducing the dimensionality of the data

- a. P and Q
- b. Q and S
- c. R and S.....
- d. P and S

Question 29.

Which of the following are correct for properties for PCA.

- a. The output of PCA are these principal components, the number of which is more than or equal to the number of original variables.

- b. The Principal Components are non-linear combinations of the original variables
- c. The variation present in the PCs decreases as we move from the 1st PC to the last one, hence the importance.....
- d. None of the above

Question 30.

Let's say you have the following distribution of PCs and their explained variance

Principal Component	Explained Variance
PC1	31%
PC2	22%
PC3	17%
PC4	8%
PC5	7%
PC6	6%
PC7	5%
PC8	4%

How many PCs would be sufficient to explain at least 90% of the variance in the dataset?

- a. 4
- b. 6.....
- c. 7
- d. 5

Question 31.

Variance is not suitable for use as a homogeneity measure for classification problems in a decision tree, because

- a. It will not give good results
- b. Variance can be computed only for real valued labels

- c. Class labels are not numeric
- d. Class labels are usually unordered and there is no 'distance' defined between classes

Question 32.

For dataset with entropy e , there happens to be an attribute A such that for any value of the label i , one can find a value j for A with $\Pr(\text{label} = i \mid \text{Attribute } A = j) = 1$. The information gain after a split on A for this dataset will be

- a. 0
- b. 0.5
- c. e
- d. Cannot Say

Question 33.

Given a dataset in which the label column is almost completely determined by a subset of k attributes, the depth of the decision tree built on this dataset is expected to be:

- a. K
- b. $\log k$
- c. k^2
- d. None of the above

Question 34.

The two most important trick that contribute to the success of the random forest algorithm are:

- a. Bootstrap Sampling and Random Feature Selection
- b. Bootstrap Sampling and Aggregation
- c. Generating lots of random trees and allowing each tree to overfit
- d. Random Feature Selection and Aggregation

Question 35.

In an election candidates are competing against each other and people are voting for one of the candidates. Voters communicate with each other while casting their votes. If you consider each voter as an individual model, which of the below techniques does this election procedure represent?

- a. Bagging
- b. Boosting
- c. Both Bagging and Boosting
- d. None of the above

Question 36.

Why would we use a random forest instead of a decision tree?

- a. For a model that is easier for a human to interpret
- b. To reduce the variance of the model
- c. For lower training error.
- d. To increase the variance of the model

Question 37.

If you use a random number generator to predict the output 0 or 1 for a binary classification problem, what will be the area under the curve of the ROC curve?

- a. 0
- b. 0.5
- c. 1
- d. 100

Question 38.

You have built a Logistic Regression model that is trying to predict whether a loan is approved or not based on a person's FICO score. Here are the model parameters: Intercept (B_0) = -9.346 and coefficient of FICO score = 0.0146. Given these parameters, can you calculate the probability of a loan getting approved for someone with a FICO score of 655?

Question 39.

Which of the following statements is NOT true?

- a. Each time the clusters are made during the K-means algorithm, the centroid is updated
- b. The cluster centers that are computed in the K-means algorithm are given by centroid value of the cluster points
- c. Standardization of the data is not important before applying Euclidean distance as a measure of similarity/dissimilarity.
- d. The centroid of a column with data points 25, 32, 34 and 23 is 28.5

Question 40.

Consider the following two Statements:

Statement 1: The distance between 2 clusters is the maximum distance between any 2 points in the clusters in complete linkage.

Statement 2: Most of the time Complete linkage will produce unstructured dendograms.

- a. Statement 1 is correct and statement 2 is wrong
- b. Statement 1 is wrong and statement 2 is correct
- c. Both statements are correct
- d. Both statements are wrong

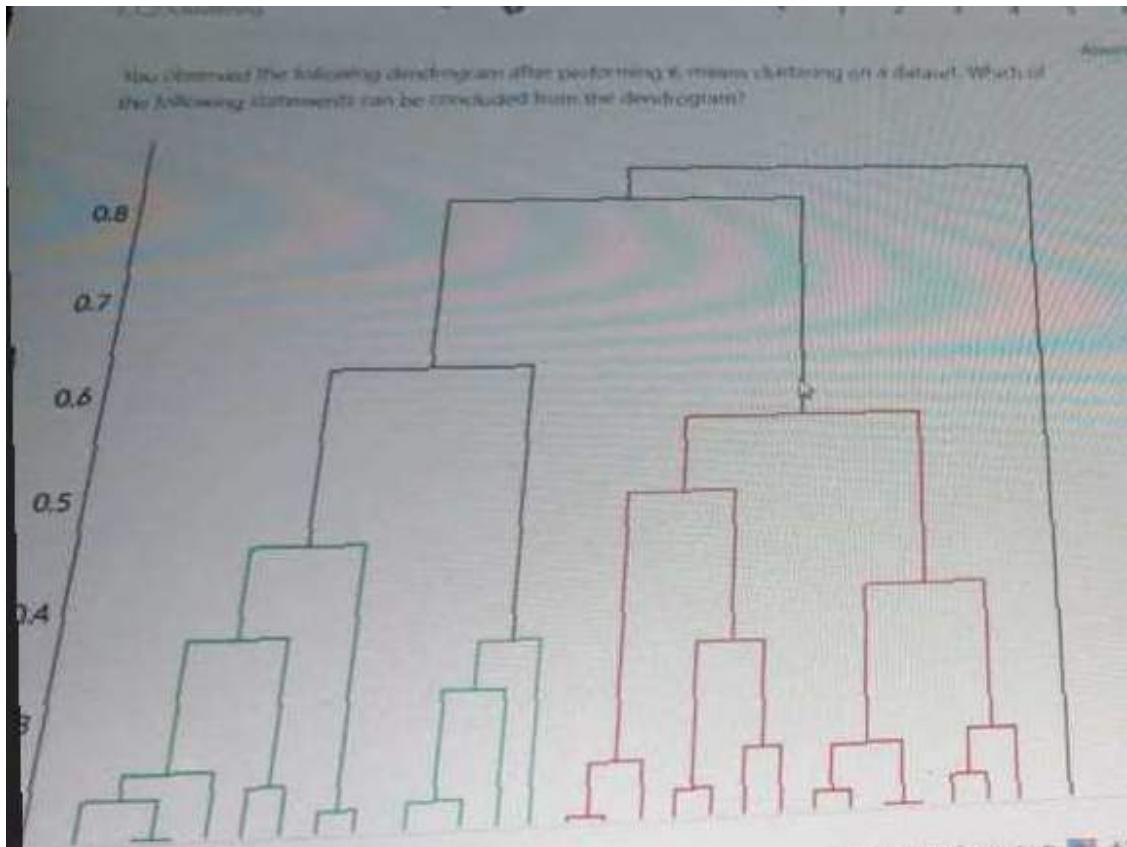
Question 41.

A client Isas approached you for a problem statement that requires the use of clustering. You decided to model the problem statement with hierarchical clustering. Consider the datasets having 'n' data points. Which of the following statements is true for the above problem statement?

- a. $n \times n$ distance matrix should be calculated the mention problem statement
- b. Initially 'n' clusters are formed for the mentioned problems statement
- c. The output for the problem statement above is a dendogram
- d. All the above

Question 42.

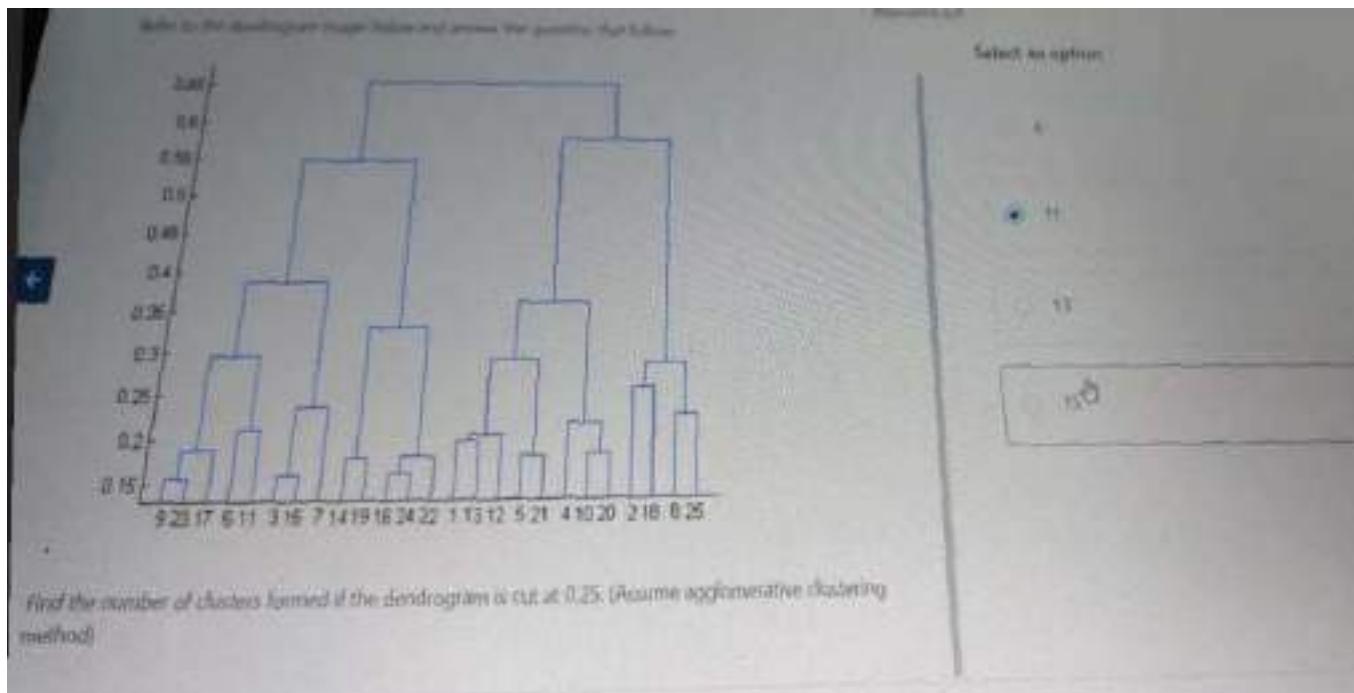
You observed the following dendrogram after performing K means clustering on a dataset. Which of the following statements can be concluded from the dendrogram?



- a. The initial number of clusters is 6
- b. There are 25 data points used in the above clustering algorithm
- c. Single linkage is used to define the distance between two clusters in the above dendrogram
- d. The above dendrogram interpretation is not possible for K-Means clustering

Question 43.

Refer to the dendrogram image below and answer the question that follow



Find the number of clusters formed if the dendrogram is cut at 0.25.(Assume agglomerative clustering method)

- a. 6
- b. 11
- c. 13
- d. 15

Question 44.

ROC curve shows the tradeoff between the True Positive Rate (TPR) and the False Positive Rate (FPR). TPR and FPR are sensitivity and (1-specificity) respectively. The following function is written in Python using metrics package from the scikit-learn library for a ROC curve function.

```
def draw_roc(actual, probs):
```

```

fpr tpr, thresholds=metrics.roc_curve(actual
probs, drop_intermediate=False)
auc_score = metrics.roc_auc_score(actual, probs)
return None

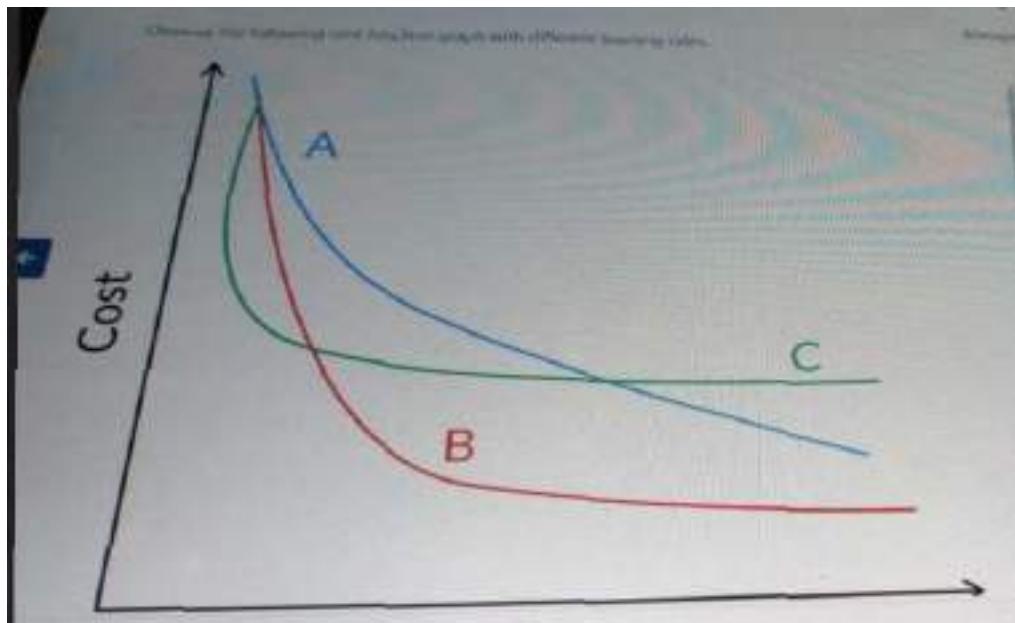
```

Which of the following statements are true? (More than one option may be correct.)

- The area under the ROC curve can be more than 1
- The arguments passed in the above function are actual values of the target variables and the predicted values (i.e. 0 or 1)
- Larger the area under the curve, the better will be the model
- The arguments passed in the above function are actual values of the target variable and the respective predicted probabilities

Question 45.

Observe the following cost function graph with different learning rates (MCQ)



- The learning rate of curve C is highest among all curves
- The learning rate for curve B is lower than A
- The learning rate for curve B is higher than A

- d. The learning rate of curve C is the smallest among all curves
- e. None of the above.

Question 46.

Ridge can be interpreted as least-square linear regression where

- a. Weights are regularized with L1 norm
- b. The weights have a Gaussian prior
- c. Weights are regularized with L2 norm
- d. The solution algorithm is simpler

Question 47.

The Zipf's Laws help us form the basic intuition for stopwords. Which of these are valid reasons for stopwords to be removed from a text.

- a. Stopwords provide no useful information especially in applications like search engines
- b. Removing stopwords reduces the size of the dataset and results in faster computation on the text data.
- c. Removing stopwords improves performance of problem, like sentiment analysis
- d. Both a and b.....
- e. Both a and c

Question 48.

In linguistic morphology _____ is the process for reducing inflected words to their root form

- a. Rooting
- b. Stemming.....
- c. Text-proofing
- d. Both a and b

Question 49.

You have a string “fellowship” that is misspelled as “feloship”. You are using the Levenshtein algorithm to calculate the edit distance for these words.

What will the edit distance for the given word be?

- a. 2.....
- b. 3
- c. 4
- d. 1

Question 50.

You have a problem statement to build a multivariate logistic regression model. There are two features say "Infected" and "Blood Group" of your interest in the dataset. The feature "Infected" takes two values "yes" or "no" whereas "Blood Group" takes multiple levels like A, A+, O, O+

Now consider the following statements.

Statement 1: For the feature "Infected", mapped in preferred over the creation of dummy variables

Statement 2: For the feature "Blood Group", the creation of dummy variables is preferred over mapping

- a. Statement 1 is correct and statement 2 is wrong
- b. Statement 2 is correct and statement 1 is wrong
- c. Both the statements are correct
- d. Both the statements are incorrect

Question 51.

Which of the following is true for weight of evidence (WoE) analysis?

- a. It helps in finding the different predictive patterns for the different segments that might be present in the data.
- b. WoE helps in treating missing values for both continuous and categorical variables

- c. WoE values should follow an increasing or decreasing trend across bins.
- d. All of the above

Question 52.

Which of the following is correct for a logistic regression model?

- a. The independent variables should not be multicollinear
- b. The dependent variable should follow Normal Distribution.
- c. The log odds in a logistic regression model lies between 0 and 1
- d. F1-score is always the best metric for evaluating a loquitie regression model.

Question 53.

The coronavirus disease (COVID-19), was declared a pandemic by the World Health Organization (WHO) in February 2020. Currently, there are no vaccines or treatments that have been officially approved by WHO after clinical trials. India has not seen the peak of infection yet and the number of infections is touching a new height daily. The business unit of an Indian health and hygiene company approaches you to know "**Why the sales of masks is decreasing despite the number of corona infections increasing daily**".

Answer the below question:

Consider the following two statements

Statement 1: Understanding the change in customer behavior is an important factor to be considered for business understanding for the problem statement above

Statement 2: One of the possible hypotheses for the above problem statement: There is a rise in the number of companies manufacturing normal/surgical masks due to which the sales of the client's company is decreasing

- a. Statement 1 is correct and statement 2 is wrong
- b. Statement 1 is wrong and statement 2 is correct
- c. Both statements are correct
- d. Both statements are wrong

Question 54.

Which of the following is NOT a methodology by which you can identify the optimal number of clusters for K-means clustering? (More than one option may be correct)

- a. Dendrogram inspection method
- b. Elbow Method
- c. Single Linkage Method
- d. Silhouette score

Question 55.

What is the advantage of using measures such as Cp, AIC BIC Adjusted R2 or mean cross-validated error?

- a. They penalise the model for being too complex
- b. They penalise the model for being too simple
- c. Both a and b
- d. None of the above

Question 56.

In a simple linear regression model when you fit a straight line through the data you'll get the two parameters of the straight line,i.e the intercept B_0 , and the slope B_1 . Which of the following is true for B_0 and B_1 ? (More than one option may be correct)

- a. The null hypothesis for a simple linear regression model is $H_0: B_1 = 0$
- b. If the p-value turns out to be greater than 0.05 for B_1 , it means B_1 is significant

- c. If B_1 turns out to be insignificant, that means there is no relationship between the dependent and independent variable
- d. If the p-value turns out to be less than 0.05 for B_0 , it means that B_0 is a non-zero

Question 57.

As the number of training examples goes to infinity, your model trained on that data will have

- a. Low Variance
- b. High Variance
- c. Same Variance

Question 58.

Consider the following statements.

- A. The principal components are linear combinations of the original variables
- B. All principal components are orthogonal to each other
- C. PCA won't work well in a dataset that is highly correlated

Which of the above statement(s) are NOT TRUE?

- a. Only A
- b. Only C
- c. A and B
- d. A, B and C

Question 59.

Recall the telecom churn example of the log odds for churn are equal to 0 for a customer, then that means

- a. There is no chance of the customer churning
- b. The probability of the customer churning is equal to the probability of the customer not churning

- c. The probability of the customer churning is very small compared to the probability of the customer not churning
- d. The probability of the customer churning is very large compared to the customer not churning

Question 60.

INPUT TABLE

You're given a **orders** table and the columns in the orders table are shown below:

orders
Order_Id INT
Type VARCHAR(10)
Red_Shipping_Days INT
Scheduled_Shipping_Days INT
Customer_Id INT
Order_Cty VARCHAR(20)
Order_Date DATE
Order_Region VARCHAR(15)
Order_State VARCHAR(20)
Order_Status VARCHAR(20)
Shipping_Mode VARCHAR(20)
Indexes

QUERY

- Calculate count of all the orders
 - ❖ where the Order State is Gujarat
 - ❖ where the Order Status is PENDING.
 - Note - Use the alias of oc for count of orders
- Group the results by Order_City
- Order them by oc & Order_City in ascending order

OUTPUT COLUMNS

oc, Order_City

here's an image showing how a sample output would look like:

OC	ORDER_CITY
1	Jamnagar
1	Rajkot
1	Vadodara
2	Jodhpur
3	Bhavnagar
4	Surat

Solution 60.

```
Select count (*) as oc, order_city  
from orders where order_state = "Gujarat "  
and order_status = "PENDING" group by order_city  
Order by oc, order_city;
```

Question 61.

INPUT TABLE

You're given a **orders** table and the columns in the orders table are shown below:

orders
Order_Id INT
Type VARCHAR(10)
Red_Shipping_Days INT
Scheduled_Shipping_Days INT
Customer_Id INT
Order_Cty VARCHAR(20)
Order_Date DATE
Order_Region VARCHAR(15)
Order_State VARCHAR(20)
Order_Status VARCHAR(20)
Shipping_Mode VARCHAR(20)

Indexes

QUERY

- Calculate count of all the orders
 - ❖ where the *Order_State* is Maharashtra
 - Note - Use the alias of oc for the count of orders
- Group the results by *Type*
- Order them by oc in ascending order

OUTPUT COLUMNS

Oc, Type

here's an image showing how a sample output would look like:

OC	TYPE
45	CASH
89	PAYMENT
108	TRANSFER
151	DEBIT

Solution61.

```
Select count (*) as oc, type [  
from orders  
where order_state = 'Maharashtra'  
group by Type  
order by oc;
```

Question 62.

How to attempt?

Given a single positive integer 'n' greater than 2, print the following pattern with all zeroes and ones such that the ones make a shape like 'Z'.

Examples:

Input 1:

3

Pattern 1:

1 1 1
0 1 0
1 1 1

Output 1:

[1,1,1,0,1,0,1,1,1]

Explanation:

The pattern is converted to list, the first row is added to the list first, then second row and then third and so on.

Input 2:

5

Pattern 2:

1 1 1 1 1
0 0 0 1 0
0 0 1 0 0
0 1 0 0 0
1 1 1 1 1

Output 2: [1, 1, 1, 1, 0, 0, 0, 1, 0 , 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 1, 1, 1, 1]

Solution62.

```
import numpy as np
```

```
n = int(input("Enter size:"))
```

```
data = np.zeros([n,n], dtype=int) # fill grid of nxn with zeros.  
data[0] = data[n-1] = np.ones(n, dtype=int) # set first and last row to 1s.
```

```
j = n-2  
for i in range(1,n-1): # fills diagonally 1s.  
    data[i][j] = 1  
    j-=1
```

```
print(data)
```

SCQs [Paper-I]

Linear Regression

1. Feature engineering is an important step in any model building exercise. It is the process of creating new features from a given data set using the domain knowledge to leverage the predictive power of a machine learning model. Which of the following statements are correct?

Statement 1: Feature engineering techniques are applied before train test split.

Statement 2: There is no difference between standardization and normalization,

Statement 3: Mean encoding is a feature engineering technique for handling categorical features.

- a. Only 1 and 2
- c. Only 2 and 3
- b. **Only 1**
- d. Only 3

2. VIF is used to detect Multicollinearity. Which of the following statements is NOT true for VIF?

- a. **The VIF has lowest bound of 0**
- b. The VIF has no upper bound
- c. VIF for a variable generally changes if you drop one of the predictor variables
- d. If a variable is a product of two other variables, it can have a high VIF

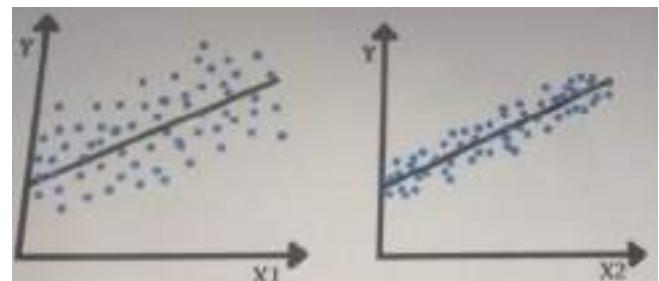
3. The distribution of errors terms in a linear regression model should look like (the horizontal line represents $y=0$):

- a. A
- b. **C**
- c. B
- d. D



4. For the same dependent variable Y, two models were created using the independent variables X1 and X2. The following graph represent the fitted line on the scatterplot. (Both the graph are on same scale). Which of the following is true about the residuals in these two models?

- a. The sum of residuals in model 2 is higher than model 1
- b. **The sum of residuals in model 1 is higher than model 2**
- c. Both have the same sum of residuals
- d. Nothing can be said about the sum of residuals from the given graph



5. You built a simple linear regression model on a provided problem statement by the client. After a few days, the client asks you to build a new model with an increased number of data points (old dataset + new data points). The count of new data points exceeds old data points by 20%.

Which of the following statement is TRUE regarding the mean of residuals?

- a. Mean of residuals of old model > Mean of residuals of new model
- b. Mean of residuals of old model < Mean of residuals of new model
- c. Mean of residuals of old model = Mean of residuals of new model
- d. **Information provided is not enough to comment on the mean of residuals**

6. A scatterplot was plotted for two variables – age and income to find out how the income depends on the age of a person. It was found that as the income increases linearly with age, the variability in income also increases. This is a violation of which of the following assumptions of linear regression?

- a. Homogeneity
- b. **Homoscedasticity**
- c. Heterogeneity
- d. Linearity

7. RFE method is used for:

- a. Dummy variable creation
- b. **Feature selection**
- c. Detecting multicollinearity
- d. Univariate regression

8. Which of the following assumptions do we make while building a simple linear regression model (assume X and y to be independent and dependent variables respectively)
- A. There is a linear relationship between X and y
 - B. X and Y are normally distributed
 - C. Error terms are independent of each other
 - D. Error terms have constant variance
- a. A, B, C and D
 - c. A, C and D
 - b. A, B and C
 - d. B, C and D
9. A client approached you with a problem statement. You decided to build a multiple linear regression model on the dataset provided. The dataset consisted of 40 features. Obviously, all features will not be significant. Selecting the relevant features manually will be a tougher task. You can use RFE to select relevant features. RFE is an automated feature selection technique. Initially, you assumed 25 features can explain your whole data. Which of the following commands correctly calls the RFE technique in Python? (Here "lm" is the fitted instance of multiple linear regression model)
- a. from stastmodel.feature_selection import RFE
rfe=RFE(lm,25)
rfe=rfe.fit(X_train,y_train)
 - b. from sklearn.feature_selection import RFE
rfe=RFE(lm,25)
rfe=rfe.predict(X_train,y_train)
 - c. from sklearn.feature_selection import RFE
rfe=RFE(lm,25)
rfe=rfe.fit(X_train,y_train)
 - d. from RFE import feature_selection
rfe=RFE(lm,25)
rfe=rfe.predict(X_train,y_train)
10. Suppose that on adding a new predictor variable to a linear regression model (model-1), the adjusted r-squared of the new model (model-2) decreases. Choose the correct statement:
- a. The r-squared of model-2 will be less than that of model 1
 - b. The r-squared of model-2 increases, but the complexity of model-2 also increases
 - c. The r-squared of model-2 decreases, but the complexity of model-2 also increases
 - d. Nothing can be said about the r-squared of model-2
11. Some of the independent variables (predictors) might be interrelated, due to which the presence of a particular independent variable in the model is redundant. This phenomenon is called Multicollinearity. Suppose that you are building a multiple linear regression model for a given problem statement, which of the following statements is TRUE w.r.t. multicollinearity?
- a. Multicollinearity is a problem when your only goal is to predict the independent variable from the set of dependent variables
 - b. Multicollinearity is a problem when your goal is to infer the effect on the dependent variable due to independent variable.
 - c. Multicollinearity is not a problem if a variable is not collinear with your variable of interest
 - d. Multicollinearity is not a problem if there are multiple dummy(binary) variables that represent a categorical variable with three or more categories
12. If the co-efficient of determination is 0.47 between a dependent variable and an independent variable. This denotes that-
- a. The relationship between the two variables is not strong
 - b. The corelation coefficient between the two variables is also 0.47
 - c. 47% of the variance in the independent variable is explained by the dependent variable
 - d. 47% of the variance in the dependent variable is explained by the independent variable
13. While solving linear regression, the dependent variable is-
- a. Numeric
 - c. Categorical
 - b. Dummy coded
 - d. Binary

14. Consider the following two assumptions for a single regression model. (Assume X and y to be independent and dependent variables respectively).

Statement 1: There is a linear relationship between X and y

Statement 2: X and y are normally distributed

- a. Statement 1 is correct and statement 2 is wrong
- b. Statement 2 is correct and statement 1 is wrong
- c. Both the statements are correct
- d. Both the statements are incorrect

15. What does standardized scaling do?

- a. Bring all data points in the range 0 to 1
- b. Bring all data points in the range -1 to 1
- c. Bring all the data points in a normal distribution with mean 0 and standard deviation 1
- d. Bring all the data points in a normal distribution with mean 1 and standard deviation 0

16. In the linear regression, F-statistic is used to determine-

- a. The significance of the individual beta coefficient
- b. The variance explanation strength of the model
- c. The significance of the overall model fit
- d. Both A and C

17. Suppose you run a regression with one of the feature variable T, with all the remaining feature variables. The R-squared of this model was found out to be 0.8. What will be the VIF for the variable T?

- a. 1.56
- c. 2.77
- b. 3.33
- d. 5.00

18. Which of the following is true regarding the error terms in linear regression?

- a. The sum of residuals should be zero
- b. The sum of residuals should be lesser than zero
- c. The sum of residuals should be greater than zero
- d. There is no such restriction on what the sum of residuals should be

Logistic Regression and Classification

19. Suppose an imbalanced data set has a class ratio of 2:3, and you want to run a cross-validation scheme to evaluate a model's performance. If you apply a stratified k-fold to generate the train-test folds, what will be the distribution of the classes in the test split?

- a. 1:5
- c. 2:3
- b. 1:7
- d. None of these

20. Consider the following two statements-

Statement 1: Suppose the value of Precision and Recall for a model is 0.65 and 0.75 respectively. Then the value of F1-score will be -0.696

Statement 2: Mean squared error is a metric that can be used to evaluate logistic regression model.

- a. Statement 1 is correct and statement 2 is wrong
- b. Statement 2 is correct and statement 1 is wrong
- c. Both the statements are correct
- d. Both the statements are incorrect

21. The output of logistic model is-

- a. 0 or 1
- c. Any value between 0 and 1
- b. 0.5
- d. Depends on the business problem

22. What is the use of performing segmentation on a dataset before running a logistic regression on it?

- a. It helps in capturing the seasonal fluctuations that might be present in the data
- b. It helps to find the optimal cut-off point more easily
- c. It helps in finding the different predictive patterns for the different set of data points that might be present in the data
- d. It helps capture the trends easily when there is a class imbalance

23. Given an imbalanced dataset, the ratio of positive to negative class is 1: 10000. You run a logistic regression model and find out the model has a high value of precision and a low value of recall. Which of the following statements is true?

- a. The class is handled well by the data
- b. The model is not able to detect the class, but when it does it is highly trustable
- c. The model is able to detect the class but it includes data points from the other class as well
- d. The class is handled poorly by the data

24. You have to build a logistic regression model that is trying to predict whether loan is approved or not based on a person's FICO score. Here are the model parameters: Intercept(β_0)= -9.346 and the co-efficient of FICO score=0.0146. Given the parameters, can you calculate the probability of loan getting approved for someone with a FICO score of 640?

- a. 0.35
- b. 0.45
- c. 0.40
- d. 0.50

25. Which of the following is correct for a logistic regression model?

- a. The independent variable should not be multicollinear
- b. The dependent variable should follow Normal Distribution
- c. The log odds in a logistic regression model lies between 0 and 1
- d. F1 score is always the best metric for evaluating a logistic regression model

26. You have a problem statement to build a multivariate logistic regression model. There are two features say 'infected' and 'Blood Group' of your interest in the dataset. The feature 'infected' takes two values "yes" or "no" whereas "Blood Group" takes multiple levels like A, A+, O, O+ etc.

Now consider the following statements-

Statement 1: For the feature "infected", mapping is preferred over the creation of dummy variables.

Statement 2: For the feature "Blood Group", the creation of dummy variables is preferred over mapping.

- a. Statement 1 is correct and statement 2 is wrong
- b. Statement 2 is correct and statement 1 is wrong
- c. Both the statements are correct
- d. Both the statements are incorrect

27. For a completely random binary classification model, what will be the area under the curve of the ROC graph?

- a. 0
- b. 0.5
- c. 0.25
- d. 1

28. Consider the following univariate logistic model

$$y = \beta_0 + \beta_1 x_1$$

Which of the following statement is NOT true?

- a. The maximum likelihood estimation determines the best combination of β_0 and β_1
- b. If β_1 is increased by 1 unit, Y increases by 1 unit
- c. β_0 is the y-intercept
- d. If β_1 is increased by 1 unit, log odds increases by 1 unit

29. You have to build a logistic regression model that is trying to predict whether loan is approved or not based on a person's FICO score. Here are the model parameters: Intercept(β_0)= -9.346 and the co-efficient of FICO score=0.0146. Given the parameters, can you calculate the probability of loan getting approved for someone with a FICO score of 655?

- a. 0.35
- b. 0.55
- c. 0.45
- d. 0.65

30. Consider the following confusion matrix. Which among the following is the lowest for the given confusion matrix?

Total=500	Actual Positive	Actual Negative
Predicted Positive	196	20
Predicted Negative	28	256

- a. Accuracy
- b. Sensitivity
- c. Precision
- d. Specificity

31. If you use a random number generator to predict the output 0 or 1 for a binary classification problem, what will be the area under the curve of the ROC curve?

- a. 0
- c. 0.5
- b. 1
- d. 100

32. How is regression different from classification?

- a. One is supervised while the other is unsupervised
- b. One is iterative while the other is closed
- c. In regression, the response variable is numeric while it is categorical in classification
- d. None of the above

33. Recall the telecom churn example. If the log odds for churn are equal to 0 for a customer, then that means-

- a. There is no chance of the customer churning
- b. The probability of customer churning is equal to the probability of the customer not churning.
- c. The probability of customer churning is very small compared to the probability of the customer not churning.
- d. The probability of customer churning is very large compared to the probability of the customer not churning.

34. Recall the telecom churn example. If the log odds for churn are equal to 1/3 for a customer, then that means-

- a. The probability of customer not churning is 3 times the probability of the customer churning
- b. The probability of customer churning is 3 times more than the probability of the customer not churning
- c. The probability of customer not churning is 4 times the probability of the customer churning
- d. The probability of customer churning is 4 times more than the probability of the customer not churning

35. Which of the following statements is NOT true?

- a. In the case of a fair coin, the odds of getting heads is 1
- b. The error values of linear and logistic regression have to be normally distributed
- c. Specificity decreases with the increase in sensitivity
- d. As TPR increases, FPR also increases

36. Take a look at the following three problem statements.

Problem statement 1: Let's say that you are building a telecom churn prediction model with the business objective that your company wants to implement an aggressive customer retention campaign to retain the high churn-risk' customers. This is because a competitor has launched extremely low-cost mobile plans, and you want to avoid churn as much as possible by incentivising the customers. Assume that budget is not a constraint.

Problem statement 2: Let's say you are building a cancer detection model with the objective that both the patient who has cancer and the patient who has not cancer can be detected correctly. It can have serious implications if you predict either of the class wrong, ie., if wrongly detected as "not cancer" the patient will die of cancer, and if wrongly detected as "cancer" the patient will die of chemotherapy.

Problem statement 3: You have to build an image classification model where 60% of images belong to one class and rest 40% images belong to another class. You have to predict the class of a new image.

Which is the correctly matched model evaluation metric for the above classification models?

- a. Problem Statement 1: Specificity
- c. Problem Statement 2: Sensitivity
- b. Problem Statement 2: Specificity
- d. Problem Statement 3: Accuracy

37. What will be the accuracy percentage of the given confusion matrix of the three-class classification?

True/Predicted	Class A	Class B	Class C
Class A	13	0	5
Class B	0	4	8
Class C	1	1	9

- a. 63%
- c. 36%
- b. 71%
- d. 45%

Clustering

38. In hierarchical clustering, the shortest distance and the maximum distance between points in two clusters are defined as and respectively.

- a. Single linkage and complete linkage
- c. Complete linkage and single linkage
- b. Single linkage and average linkage
- d. Complete linkage and average linkage

39. Which of the following statement is NOT true?

- a. Each time the clusters are made during the K-means algorithm, the centroid is updated.
- b. The cluster centres that are computed in the K-means algorithm are given by centroid value of the cluster points
- c. Standardization of the data is not important before applying Euclidean distance as a measure of similarity/dissimilarity
- d. The centroid of a column with data points 25, 32, 34 and 23 is 28.5.
- e. The Euclidean distance between two points (10,2) and (4,5) is 7.

40. Initializing the following command in Python will result in the following:

```
model_clus= KMeans(n_clusters=6, max_iter=50)
```

- a. Run maximum 6 iterations
- c. Run maximum 40 iterations
- b. Create 6 final clusters
- d. Create 50 final clusters

41. Which of the following is not true for Hopkins Statistics?

- a. Hopkins statistics decides if the data is suitable for clustering or not
- b. Hopkins statistics lie between -1 and 1
- c. If the Hopkins statistics comes out to be 0, then the data is uniformly distributed
- d. If the Hopkins statistics comes out to be 1, then the data is highly suitable for clustering

42. Consider the two statements-

Statement 1: The distance between 2 clusters is the maximum distance between 2 points in the clusters in complete linkage.

Statement 2: Most of the time Complete linkage will produce unstructured dendograms.

- a. Statement 1 is correct and statement 2 is wrong
- b. Statement 2 is correct and statement 1 is wrong
- c. Both the statements are correct
- d. Both the statements are incorrect

43. A client has approached you for a problem statement that requires the use of clustering. You decided to model the problem statement with hierarchical clustering. Consider the datasets having 'n' data points.

Which of the following statements is true for the above problem statement?

- a. 'n*n' distance matrix should be calculated for the mentioned problem statement
- b. Initially 'n' clusters are formed for the mentioned problem statement
- c. The output of the problem statement above is a dendrogram
- d. All the above

44. Silhouette metric for any ith point is given by $S(i) = (b(i) - a(i))/\max(a(i), b(i))$

Which of the following is not true about the Silhouette metric?

- a. $b(i)$ is the average distance from the nearest neighbour cluster (Separation)
- b. $a(i)$ is the average distance from own cluster (Cohesion).
- c. If $S(i) = 1$ the data point is similar to its own cluster.
- d. Silhouette metric ranges from 0 to +1

45. Clustering is used to identify the below-

- a. Data distribution
- b. Principal components
- c. Correlation among the data points
- d. Subgroups in the data

46. For a K-means clustering process, the Hopkin Statistic for the dataset came out to be 0.8. Hence the dataset is-

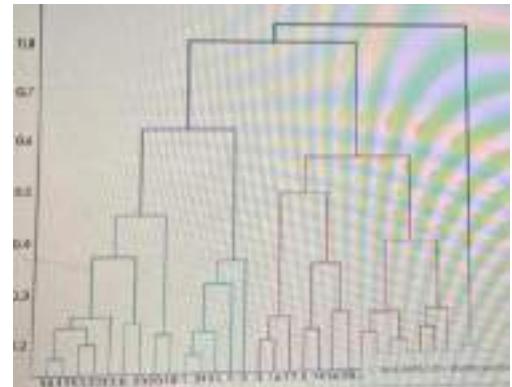
- a. Suitable for clustering
- b. Can't say from the given information
- c. Not suitable for clustering
- d. None of the above

47. For a K-means clustering process, the Hopkin Statistic for the dataset came out to be 0.3. Hence the dataset is-

- a. Suitable for clustering
- b. Can't say from the given information
- c. Not suitable for clustering
- d. None of the above

48. You observed the following dendrogram after performing K-means clustering on a dataset. Which of the following statements can be concluded from this dendrogram?

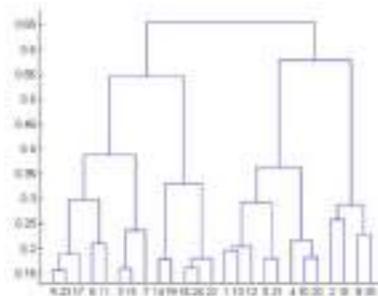
- a. The initial number of clusters is 6
- b. There are 25 data points used in the above clustering algorithm
- c. Single linkage is used to define the distance between two clusters in the above dendrogram
- d. The above dendrogram interpretation is not possible for K-means clustering.



49. Refer to the dendrogram image below and answer the question that follow:

Find the number of clusters formed if the dendrogram is cut at 0.25. (Assume agglomerative clustering method)

- a. 6
- b. 13
- c. 11
- d. 15



Decision Tree

50. Which of the following is the correct sampling technique that is used by a random forest model to overcome the problem of overfitting?

- a. Random sampling
- b. Oversampling
- c. Bootstrapping
- d. Stratified sampling

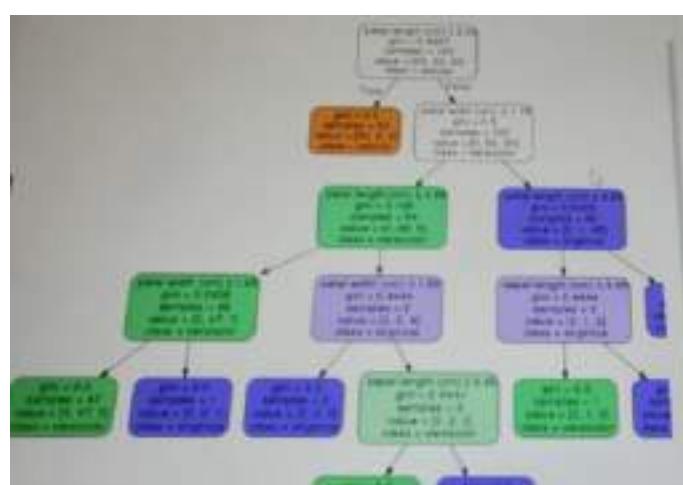
51. Which of the following metrics measures how often a randomly chosen element would be incorrectly identified?

- a. Entropy
- b. Gini Index
- c. Information Gain
- d. None of these

52. Which of the following is true for weight of evidence (WoE) analysis?

- a. It helps in finding the different predictive patterns for the different segments that might be present in the data
- b. WoE helps in treating missing values for both continuous and categorical variables
- c. WoE values should follow an increasing or decreasing trend across bins.
- d. All of the above

53. Refer to the decision tree given below and choose the statement that is correct as per this tree.



- a. The tree given above will show very good performance on the train data
- b. The tree given above is an underfitting tree.
- c. If the petal length is more than 2.45, then it is equally likely that the flower is either setosa or virginica.
- d. Both B and C

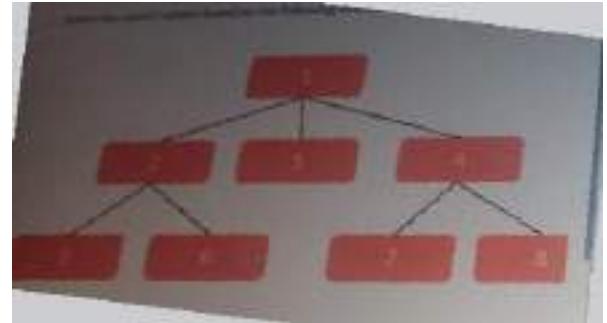
54. Suppose you train a decision tree with the following data. Which feature should we split on at the root?

X	Y	Z	V
T	T	F	1
F	F	F	0
T	T	T	0
F	T	T	1

- a. X
b. Z
c. Y
d. Cannot be determined

55. Select the correct option based on the following decision tree.

- I. Node 8 is the root node
II. Leaf node is 5
III. Nodes 2, 3, 4 are internal nodes.
IV.
a. Only I
b. Both II and III
c. Only II
d. Both II and IV



NLP and Lexical Processing

56. Choose the correct option from the following:

The difference between “+” and “*” quantifier is-

- a. ‘+’ needs the preceding character to be present at least once whereas ‘*’ does not need the same.
b. ‘*’ need the character to be present at least once whereas ‘+’ does not need the same.
c. Both then quantifiers have same functionality
d. None of the above

57. What is the Levenshtein distance between ‘decade’ and ‘dictate’?

- a. 3
b. 5
c. 4
d. 6

58. Which of the following strings will match the expression ‘^01+0\$’?

1. 0
2. 00
3. 011110
a. Only option 1
b. Both 1 and 2
c. Only option 3
d. Both 2 and 3

59. What is the Levenshtein distance between ‘shutter’ and ‘shelter’?

- a. 1
b. 3
c. 2
d. 4

60. Which of the following strings will match with the regular expression ‘^01*0\$’?

1. 0
2. 00
3. 0111111110
a. Only option 1
b. Both 1 and 2
c. Only option 3
d. Both 2 and 3

Business Problem Solving

61. The coronavirus disease (COVID-19) was declared a pandemic by World Health Organization (WHO) in February 2020. Currently, there are no vaccines or treatments that have been officially approved by WHO after clinical trials. India has not seen the peak of infection yet and the number of infections is touching a new height daily. The business unit of an Indian health and hygiene company approaches you to know "**Why the sales of masks is decreasing despite the number of corona infections increasing daily**".

Answer the following questions:

Suppose you mapped the above problem statement with a classification problem, either a customer will buy a mask or not. You'll build model as your initial solution.

- a. Neural Network
- c. **Logistic Regression**
- b. Decision Tree
- d. All of the above

62. The coronavirus disease (COVID-19) was declared a pandemic by World Health Organization (WHO) in February 2020. Currently, there are no vaccines or treatments that have been officially approved by WHO after clinical trials. India has not seen the peak of infection yet and the number of infections is touching a new height daily. The business unit of an Indian health and hygiene company approaches you to know "**Why the sales of masks is decreasing despite the number of corona infections increasing daily**".

Answer the below questions:

Consider the following two statements:

Statement 1: Understanding the change in customer behaviour is an important factor to be considered for business understanding for the problem statement above

Statement 2: One of the possible hypotheses for the above problem statement: There is a rise in the number of companies manufacturing normal/surgical masks due to which the sales of the client's company is decreasing

- a. Statement 1 is correct and statement 2 is wrong
- b. Statement 2 is correct and statement 1 is wrong
- c. Both the statements are correct**
- d. Both the statements are incorrect

MCQs [Paper -I]

63. ROC curve shows the trade-off between the True Positive Rate (TPR) and the False Positive Rate (FPR). TPR and FPR are sensitivity and (1-Specificity) respectively. The following function is written in Python using metrics package from the sci-kit learn library for the ROC curve function.

```
def draw_roc(actual,probs):
```

```
    fpr,tpr,thresholds = metrics.roc_curve(actual, probs, drop_intermediate=False)
```

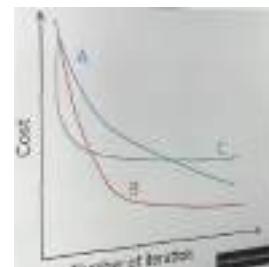
```
    auc_score = metrics.roc_auc_score(actual,probs)
```

```
    return None
```

Which of the following statements are true? (More than one option may be correct)

- a. The area under ROC curve can be more than 1
- b. The arguments passed in the above function are actual values of the target variable and the predicted values (i.e. 0 or 1)
- c. The area under the ROC can take any value between 0 and 1
- d. Larger the area under the curve, the better will be the model
- e. The arguments passed in the above function are actual values of the target variable and the respective predicted probabilities

64. Observe the following cost function graph with different learning rates.



- a. The learning rate of the Curve C is highest among all curves.
- b. The learning rate of the Curve B is lower than A.
- c. The learning rate of the Curve B is higher than A.
- d. The learning rate of Curve C is smallest among all curves.
- e. None of the above

65. Which of the following command correctly builds a logistic regression model in Python?

(More than 1 option can be correct)

- a.

```
from sklearn.linear_model import LogisticRegression  
lr=LogisticRegression()  
lr.fit(X_train,y_train)
```
- b.

```
Import statsmodel.api as sm  
lr=sm.GLM(y_train,(sm.add_constant(X_train)),  
family =sm.families.Binomial())  
lr.fit()
```
- c.

```
from sklearn.linear_model import LogisticRegression  
lr=LogisticRegression()  
lr.predict(X_train,y_train)
```
- d.

```
Import statsmodel.api as sm  
lr=sm.GLM(y_train,(sm.add_constant(X_train)),  
family =sm.families.Binomial())  
lr.predict()
```

66. In a simple linear regression model when you fit a straight line through the data you'll get the two parameters of the straight line i.e. the intercept β_0 and the slope β_1 . Which of the following is true for β_0 and β_1 ? (More than one option may be correct)

- a. The null hypothesis for a simple linear regression model is $H_0: \beta_1=0$
- b. If the p-value turns out to be greater than 0.05 for β_1 , it means β_1 is significant
- c. If β_1 turns out to be insignificant, that means there is no relationship between the dependent and the independent variable.
- d. If the p-value turns out to be less than 0.05 for β_0 it means that β_0 is non-zero

67. Which of the following metrics can be used for finding the appropriate number of clusters in K-means clustering?

(More than one option may be correct)

- a. Silhouette Score
- b. Hopkins Statistic
- c. Elbow Curve
- d. Dendrogram

68. Which of the following statements is true? (More than one option may be correct)

- a. TSS(Total Sum of Squares) is defined as the sum of all squared differences between the observed dependent variable and its mean
- b. R-Squared can take any value between 0 and 1
- c. Larger the R-squared value, the better the regression model fits the observations
- d. If RSS=5.50 and TSS=11, the value of VIF will be 1.33

69. Which of the following statements are correct in the context of logistic regression? (More than one option may be correct)

- a. The dummies for continuous variables make the model more unstable
- b. Weight of Evidence (WoE) helps in treating missing values for both continuous and categorical variables
- c. WoE should follow a non-monotonic trend across bins.
- d. Data clumping can be a problem with transforming continuous variables to dummies.
- e. Information Value or IV is an important indicator of predictive power.

70. Which of the following is NOT a methodology by which you can identify the optimal number of clusters for K-means clustering? (More than one option may be correct)

- a. Dendrogram inspection method
- b. Single Linkage method
- c. Elbow method
- d. Silhouette Score

71. Any Business Problem Solving will have the following steps:

1. To identify the right data sources, that will be useful in formulating the final solution
2. Develop hypothesis and assess the overall impact of the hypothesized solution
3. Asking the right question for business and problem understanding.
4. Define the solution approach: What will be the POC model? What will be the metrics for the model evaluation etc.
5. Converting business problem to a data science problem
6. Start your model building process with the simple POC model. And then increase the complexity of the POC model and optimize the parameters to get the best result.
7. Performing EDA on the datasets
8. Model Evaluation

What will be the correct flow for solving the above/any business problem?

- a. 3>1>5>2>4>7>6>8
- b. 3>2>1>5>4>7>6>8
- c. 4>3>1>2>5>7>6>8
- d. 3>2>1>5>4>7>8>6

Question 1

Revisit Later

Select an

In a particular game series, the following cumulative probability table has been prepared where X is the number of points scored by a particular player in one game.

x	$F(x) = P(X \leq x)$
20	0.2
40	0.26
60	0.4
80	0.8
120	0.95
200	1

If the player played 40 games in that series, calculate the number of games in which he scored more than 80 points but less than or equal to 120 points



10

12

14



6

Question 2

 Revisit Later

Select an option

For a random variable that is normally distributed the mean comes out to be 6 and the standard deviation comes out to be 1. In any given experiment, what would be the probability that the value of this random variable lies between 5 and 7? You can use the Z-tables given below.

<i>z</i>	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
-1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
-1.5	.0688	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
-1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
-1.3	.0966	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
-1.2	.1151	.1131	.1112	.1093	.1075	.1055	.1038	.1020	.1003	.0985
-1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
-1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
-0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
-0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
-0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2146
-0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
-0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
-0.4	.3446	.3403	.3372	.3330	.3290	.3254	.3220	.3182	.3158	.3121
-0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
-0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3934	.3897	.3859
-0.1	.4562	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4285	.4247
-0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

<i>z</i>	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817

98%

99.7%

95.4%

97.5%

Question 3 Revisit Later

Select an option

For a particular experiment, the following table was prepared. Here, the first column contains values of the random variable X and the second column contains the associated probabilities. However, one particular value and its associated probability are missing from the data. If the Expected Value of the random variable X came out to be $26/7$, calculate both the missing value and its associated probability.

X (values that random variable X can take)	P(X)
1	1/7
3	2/7
7	1/7

 $X=3, P=1/7$ $X=3, P=2/7$ $X=4, P=3/7$ $X=4, P=4/7$ 

Question 4 Revisit Later

Select an option

Which of the following sample sizes would result in the largest value of standard error? (Assuming that standard deviation = 2.3)

 225 64 100 49

Question 5 Revisit Later

Select an option

The dean of a college wants to know the average time spent (in hours) by the students of his college in the library. He takes a sample of 81 students visiting the library and calculates the mean time and standard deviation, which comes out to be 120 and 15, respectively. Find the interval in which the average time spent in the library by the entire population might lie with a confidence level of 95%. You can use the Z-table given below.

<i>z</i>	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817

 (112.4, 128.71) (112.57, 124.42) (116.73, 123.26) (120.74, 122.75)

Question 4 Revisit Later

Select an option

Which of the following sample sizes would result in the largest value of standard error? (Assuming that standard deviation = 2.3)

 225 64 100 49

4. Inferential Statistics & Hypothesis Testing

Attempted: 4/9

Question 6

Revisit Later

Select an option

Let's say you are conducting a hypothesis test using the p-value method. You took 2 samples A and B from the same population such that both of them have the same sample mean and the same standard deviation. If the sample size of B is 1/4th the size of sample A, how would the Z-score differ for them given the same hypothesis (statements)?

Z-score would remain the same for both of them.

Z-score for B would be 4 times that of A.

Z-score for A would be 1/2 of that of B.

Question 7 Revisit Later**Select an option**

Consider the following statements regarding the null and alternate hypothesis :

1. If a claim is made about something, then that statement automatically becomes the null hypothesis of our hypothesis test.
2. The null hypothesis and alternate hypothesis have no overlap, i.e. they are complementary to each other.
3. The null hypothesis always has the equal to, greater than, equal to or less than equal to notation associated with its statement.
4. If you fail to reject the null hypothesis, it does not mean that there is no change in the status quo, it is just that you do not have sufficient evidence to disprove it.

Which of the above statements are true?

 All of them Only 2,3,4 Only 1 and 2 None of them

Attempted: 7/9

Question 8 Revisit Later

Select an option

A supply chain that produces floor mats has been stable for a long period of time. The average weight of the floor mat is considered to be equal to 300 gms. A new analyst is hired who challenges the status quo. The analyst takes a sample of 36 floor mats with a sample mean of 310 gms and a standard deviation of 30 gms, respectively. Calculate the p-value for the test at the significance level of 5%. You can use the z-table given below:

\pm	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6199	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6594	.6591	.6628	.6664	.6701	.6738	.6772	.6808	.6844	.6879
0.5	.6995	.6990	.7019	.7054	.7088	.7123	.7157	.7190	.7224	
0.6	.7397	.7291	.7324	.7357	.7390	.7422	.7454	.7486	.7517	.7549
0.7	.7790	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7981	.7910	.7939	.7967	.7995	.8023	.8051	.8079	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8433	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621	
1.1	.8603	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8869	.8889	.8889	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9046	.9066	.9082	.9093	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9255	.9279	.9292	.9305	.9319
1.5	.9352	.9345	.9357	.9370	.9382	.9394	.9406	.9419	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9494	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9615	.9625	.9633
1.8	.9641	.9649	.9655	.9664	.9671	.9678	.9685	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9776	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817

 15.32% fail to reject the null hypothesis 15.32% reject the null hypothesis 25.02% fail to reject the null hypothesis 25.02% reject the null hypothesis

Question 9 Revisit Later**Select an option**

The daily average time spent by customers on Netflix in the Hollywood movies section is 1.5 hours. The India Head of Netflix wants to verify whether this holds true for Indian customers. Their team plans to record the daily average duration of time spent by Indian customers in the Hollywood movies section to be used as the sample to determine whether it is significantly different from 1.5 hours.

The following hypothesis is used:

$$H_0: \mu = 1.5$$

$$H_1: \mu \neq 1.5$$

Consider the following statements:

Now consider the following statements regarding the types of errors that can happen in the above hypothesis test:

Statement 1: A Type-I error happens when the India Head of Netflix concludes that the average duration spent by Indian customers in the Hollywood movies section is not 1.5 hours when it actually is.

Statement 2: A Type-I error happens when the India Head of Netflix concludes that the average duration spent by Indian customers in the Hollywood movies section is not 1.5 hours when it actually is not.

Statement 3: A Type-II error happens when the India Head of Netflix concludes that the average duration spent by Indian customers in the Hollywood movies section is 1.5 hours when it actually is.

Statement 4: A Type-II error happens when the India Head of Netflix concludes that the average duration spent by Indian customers in the Hollywood movies section is 1.5 hours when it actually is not.

Which of the above 4 statements are correct?

 1 and 4 1 and 3 2 and 3 2 and 4

Attempted: 0/4

Question 1 Revist Later

Select an option

You are an analyst working for Netflix and trying to figure out the kind of factors that affect the 'Viewer Rating' for a current crop of 25 Netflix original movies and you prepared the following correlation matrix.

	"Budget (\$M)"	"Gross_earning (\$M)"	"Movie_Length (min)"	"Viewer Rating"
"Budget (\$M)"	1			
"Gross_earning (\$M)"	0.95	1		
"Movie_Length (min)"	0.1	-0.23	1	
"Viewer Rating"	0.54	0.85	-0.33	1

Now answer the following questions.

Which of the following attributes inversely affect the 'Viewer Rating'?

- Gross earning
- Movie Length
- Both a and b
- Budget only

Attempted 1/4

Question 2

 Revisit Later

Select an option

You are an analyst working for Netflix and trying to figure out the kind of factors that affect the 'Viewer Rating' for a current crop of 25 Netflix original movies and you prepared the following correlation matrix.

	"Budget (\$M)"	"Gross Earning (\$M)"	"Movie Length (min)"	"Viewer Rating"
"Budget (\$M)"	1			
"Gross Earning (\$M)"	0.95	1		
"Movie Length (min)"	0.1	-0.23	1	
"Viewer Rating"	0.54	0.85	0.33	1



Statement 1 is correct. Statement 2 is incorrect.

Statement 1 is false. Statement 2 is true.

Both statements are correct.

Both statements are incorrect.

Consider the following two statements-

- i) 'Movie Length' and 'Gross Earning' are inversely related since they have a negative correlation value.
- ii) 'Movie Length' and 'Budget' are inversely related since they have a very low correlation value.

Given the above statements, choose the most appropriate option.

Question 3 Revisit Later

Select an option

Consider the following table and answer the questions.

Given below is the data of school boys and girls of an Indian state who reported the number of times they play games, i.e. whether they play every day, never, once a month or once a week.

	Everyday	Never	Once a month	Once a week	Total
Boys	3474	154	150	760	4558
Girls	2886	175	200	1046	4307

What is the approximate percentage of boys who play at least once a month as compared to all the students?

57%
3%
50%
32%

Question 4 Revisit Later**Select an option**

- A) Scatterplot
- B) Histogram
- C) Heatmap
- D) Both A and B

32	90	Biology
22	67	Biology
12	71	Biology
17	90	Biology
3	89	Biology

Students

Student_Name	Student_ID	Gender
Sanket Dhoble	12	Male
Aruna Vijayan	22	Female
Shashank Singh	17	Male
Sumit Rakshit	32	Male
Amit Kumar Manjhi	3	Male

Teachers

Name	Id	Age	Course_Taught
Mehul Sayani	19	24	Physics
Armit Makhija	16	35	Chemistry
Dimbeswar Rabha	7	27	Biology

Which of the following queries will display the Student_ID for only the student who has got the lowest marks in Chemistry?

Select an option

select Student_ID
from Marks
where Course = 'Chemistry'
order by Marks
limit 1;

select Student_ID
from Marks
where Course = 'Chemistry'
order by Marks;

select Student_ID
from Marks
where Course = 'Chemistry'
order by Marks desc
limit 1;

select Student_ID
from Marks
where Course = 'Chemistry'
order by Marks desc
limit 10;

		Course_Taught
32	90	Biology
22	67	Biology
12	71	Biology
17	90	Biology
3	89	Biology

Attempted: 6/7

```
- select marks
from marks
);
```

select student_name
from students
inner join marks
using (Student_ID)
where
{
select course_taught
from teachers
where name = 'mehul sayani'
) and marks = {
select max(marks)
from marks
};

select Student_Name
from Students
inner join Marks
using (Student_ID)
where Course =
{
select Course_Taught
from Teachers
where name = 'mehul sayani'
) and marks = {
select min(marks)
from marks
};

Students

Student_Name	Student_ID	Gender
Sanket Dhoble	12	Male
Aruna Vijayan	22	Female
Shashank Singh	17	Male
Sumit Rakshit	32	Male
Amit Kumar Manjhi	3	Male

Teachers

Name	Id	Age	Course_Taught
Mehul Sayani	19	24	Physics
Amit Makhija	16	35	Chemistry
Dimbleswar Rabha	7	27	Biology

Which of the following queries will display the name of the student obtaining the highest marks in the course taught by 'Mehul Sayani'?

5. SQL MCQs

« 1 2 3 4 5 6 7 » ⌂ ⌃

Attempted: 2/7

22	67	Biology
12	71	Biology
17	90	Biology
3	89	Biology

Students

Student_Name	Student_ID	Gender
Sanket Dhoble	12	Male
Aruna Vijayan	22	Female
Shashank Singh	17	Male
Sumit Rakshit	32	Male
Amit Kumar Manjhi	3	Male

Teachers

Name	Id	Age	Course_Taught
Mehul Sayani	19	24	Physics
Amit Makhija	16	35	Chemistry
Dimbeswar Rabha	7	27	Biology

Which of the following queries will print the details of the Student who scored the highest marks in physics? More than one option may be correct.

Choose the best option

 select Student_ID, Course, Student_Name
from Marks inner join Students
using(Student_ID)
where Course = 'Physics' and Marks = (select max(Marks)
from Marks);

 select Student_ID, course, student_name
from marks inner join students
using(Student_ID)
where course = 'physics' and marks = (select
max(marks)
from marks);

 select s.Student_ID, Course, Student_Name
from Marks m inner join Students s
on m.Student_ID = s.Student_ID
where Course = 'Physics' and Marks = (select
max(Marks)
from Marks);

 None of the above

Question 4 Review Later

Select an option

A CASE statement in SQL is equivalent to which of the following?

 A way to use CASE based logic in SQL A way to use IF-THEN-ELSE logic in SQL A way to define CASE while creating tables A way to use FUNCTIONS in SQL

Question 3

 Revisit Later

What will be the output of the given SQL query for the following table 'employees'?

```
1 SELECT dept_id, last_name, salary,
2 FROM employees
3 ORDER BY salary DESC;
4 GROUP BY dept_id;
5 HAVING COUNT(last_name) > 1;
```

last_name	salary	dept_id
Sutherland	54000	45
Yates	80000	45
Erickson	42000	45
Parker	57500	30
Gates	65000	30

dept_id	last_name	salary	lead_salary
45	Erickson	42000	NULL
45	Sutherland	54000	42000
30	Parker	57500	54000
30	Gates	65000	57500
45	Yates	80000	65000

dept_id	last_name	salary	lead_salary
45	Yates	80000	85000
30	Gates	65000	57500
30	Parker	57500	54000
45	Sutherland	54000	42000
45	Erickson	42000	NULL

dept_id	last_name	salary	lead_salary
45	Yates	80000	NULL
30	Gates	65000	80000
30	Parker	57500	65000
45	Sutherland	54000	57500
45	Erickson	42000	54000

Question 6

 Review later

Select an option:

Select which of the statements is true regarding user-defined functions and stored procedures.

- A user-defined function cannot call a stored procedure.
- A stored procedure cannot call a user-defined function.
- A stored procedure must return a value.
- A user-defined function supports both the input and output parameter.

Question 7

 Retake later

Select an option

For the given table 'films':

ID	release_year	rating
1	2015	8.0
2	2015	8.5
3	2015	9.0
4	2016	8.2
5	2016	8.4
6	2017	7.0

Which of the following queries will result in the given output?

ID	release_year	rating	year_avg
1	2015	8.0	8.5
2	2015	8.5	8.5
3	2015	9.0	8.5
4	2016	8.2	8.3
5	2016	8.4	8.3
6	2017	7.0	7.0

```
1 SELECT f.id, f.release_year, f.rating, AVG(rating)
2 OVER (PARTITION BY rating) AS year_avg
3 FROM films f ORDER BY release_year, rating;
```

```
1 SELECT f.id, f.release_year, f.rating, AVG(rating)
2 OVER (GROUP BY rating) AS year_avg
3 FROM films f ORDER BY release_year, rating;
```

```
1 SELECT f.id, f.release_year, AVG(rating)
2 OVER (PARTITION BY release_year) AS year_avg
3 FROM films f ORDER BY rating, release_years;
```

```
1 SELECT f.id, f.release_year, f.rating, AVG(rating)
2 OVER (PARTITION BY release_year) AS year_avg
3 FROM films f ORDER BY release_year, rating;
```

6. Python Coding

Question 1

□ Review Later

How to Attempt?

Given a dictionary containing the name and marks of some students, find out how many students have marks at least 80 using dictionary comprehension.

Example:

Import 1

[Vash Pandya]; 68, "Debaditya Basu"; 89, "Shivam Gupta"; 81, "Varun Goyal"; 77, "Niranjan Gyancha"; 54]

Output 1:

3

[Learn more about our](#)

PYTHON

Compiler: Python 3.6

```
1  # Read only region start
2  class UserMainCode(object):
3      @classmethod
4      def DlttFunc(cls, Input1):
5          ...
6
7          input1 : HashMap[String, Integer]
8
9          Expected return type : int
10         ...
11
12         # Read only region end
13         # Write code here
14
15
16
17         return(# Return your final count here)
18
```

Question 2

How to Attempt?

Given a string of the lyrics of a song, find out how many words begin with a vowel.

PS: The lyric string contains only words separated by spaces and spaces. There are no special characters like punctuation marks.

Examples:

Input 1:

Lyrics: "city of stars are you shining yet so far away
shined so brightly"

Output 1:

Question 1

How to Attempt?

INPUT TABLE

You're given a **orders** table and the columns in the orders table are shown below:

The screenshot shows a table window titled "orders". The table has 11 columns listed vertically. The first column, "Order_Id", is marked with a yellow question mark icon. The other columns are marked with a blue diamond icon. The columns are: Order_Id, Type, Real_Shipping_Days, Scheduled_Shipping_Days, Customer_Id, Order_City, Order_Date, Order_Region, Order_State, Order_Status, and Shipping_Mode. Below the table, there is a section labeled "Indexes" with a left arrow icon.

Order_Id	Type	Real_Shipping_Days	Scheduled_Shipping_Days	Customer_Id	Order_City	Order_Date	Order_Region	Order_State	Order_Status	Shipping_Mode
----------	------	--------------------	-------------------------	-------------	------------	------------	--------------	-------------	--------------	---------------

QUERY

- Calculate count of all the orders.
 - **Note** - Use the alias of **oc** for count of orders.
- **Group the** results by *Order_Date*
- **Order them** by *oc* & *Order_Date* in *ascending order*
- **Limit to 10** results.

OUTPUT COLUMNS

oc, Order_Date

2. Python for Data Science & Visualisation

Attempted 0/1

Question 1

Suppose you have three dataframes named esp1, esp2 and esp1_addn, as given below:

	Team	Played	Won	Draws	Points
0	Chelsea	38	30	8	93
1	Tottenham	38	26	8	86
2	ManCity	38	23	9	79
3	Liverpool	38	22	3	78
4	Arsenal	38	21	4	75
5	ManUtd	38	18	5	75

These three dataframes are combined in such a way that the resultant dataframe looks as follows:

	Team	Played	Won	Draws	Points
0	Chelsea	38	30	3	93
1	Tottenham	38	26	8	86
2	ManCity	38	23	9	79
3	Liverpool	38	22	10	78
4	Arsenal	38	21	6	75
5	ManUtd	38	18	15	69

Which of the following commands was used to do this?

pd.concat([esp1.append(esp2), ignore_index = False], esp1_addn), axis = 0

pd.concat([esp1.append(esp2), ignore_index = True], esp1_addn), axis = 0

pd.concat([esp1.append(esp2, ignore_index = False), esp1_addn], axis = 0)

pd.concat([esp1.append(esp2, ignore_index = True), esp1_addn], axis = 1)

Question 2 Revise Later**Select an option**

Suppose you have the following two pandas series:

```
S1 = pd.Series([0, 2, 4, 6, 8, 10, 12])  
S2 = pd.Series([0, 3, 6, 9, 12, 15])
```

What will the output of the following command be?

```
S1[S1.join(S2)[index == S2[S2.join(S1)]].index]
```

 True False [False, True, False]

A yellow hand-drawn style heart icon with a black outline, positioned next to the correct answer choice.

Question 5 [Revisit Later](#)[Select an option](#)

Consider the dataframe (df1) provided below. It has the details of the highest scores made by Indian batsmen in Test matches.

	Name	Highest_Score	Venue	Opponent
0	Rahul Dravid	270	Rawalpindi	Pakistan
1	Sunit Gavaskar	236	Chennai	West Indies
2	Sachin Tendulkar	248	Dhaka	Bangladesh
3	Saurav Ganguly	239	Bangalore	Pakistan
4	WVS Laxman	281	Kolkata	Australia

- df1.loc[[3], ["Name", "Highest_Score", "Venue", "Opponent"]]
-  df1.loc[[2], ["Name", "Highest_Score", "Venue", "Opponent"]]
- df1.loc[[1, 0, 1, 2, 3]]
- None of the above

Which of the following code snippets will fetch the row containing the data given below?

[2 Sachin Tendulkar 248 Dhaka Bangladesh]

Question 6 Revisit Later

Select an option

Suppose you have two dataframes, df1 and df2, as given below.

df1

	Name	Highest_Score	Venue	Opponent
0	Rahul Dravid	270	Rawalpindi	Pakistan
1	Sunil Gavaskar	236	Chennai	West Indies
2	Sachin Tendulkar	248	Dhaka	Bangladesh
3	Saurav Ganguly	239	Bangalore	Pakistan
4	VVS Laxman	281	Kolkata	Australia

df2

	Venue	Country
0	Rawalpindi	Pakistan
1	Chennai	India
2	Dhaka	Bangladesh
3	Bangalore	India
4	Kolkata	India

Now, these two dataframes are combined in such a way that the resultant dataframe looks like the one given below.

	Name	Highest_Score	Venue	Opponent	Country
0	Rahul Dravid	270	Rawalpindi	Pakistan	Pakistan
1	Sunil Gavaskar	236	Chennai	West Indies	India
2	Sachin Tendulkar	248	Dhaka	Bangladesh	Bangladesh
3	Saurav Ganguly	239	Bangalore	Pakistan	India
4	VVS Laxman	281	Kolkata	Australia	India

Which of the following commands was used to get the above dataframe?

 df1.append(df2) df2.merge(df1, how='inner', on = 'Country') df2.append(df1) df1.merge(df2, how='inner', on = 'Venue')

Attempted: 6/7

Question 7 Revisit Later**Select an option**

Suppose you want to roughly find out whether two numeric variables are correlated or not. Choose the most appropriate plot for this task.

 Boxplot Rug Plot Histogram Scatter Plot

Question 2

Suppose you have the following two pandas series:

`S1 = pd.Series([0, 2, 4, 6, 8, 10, 12])`

`S2 = pd.Series([0, 3, 6, 9, 12, 15])`

What will the output of the following command be?

`S1[S1.isin(S2)].index == S2[S2.isin(S1)].index`

Question 3

 Revisit Later

Select an option

Which of the following code snippets will give the output as ['u', 'p', 'G', 'Y', 'a', 'd']?

a)
def split(word):
 return list(word)

word = 'upGrad'
print(split(word))

b)
[ch for ch in 'upGrad']

c)
chars=[]
for char in char:
 chars.append(upGrad)
print(chars)

 d) Both a and b

e) None of the above

1. Python MCQs

1

Question 2

What will the output of the following program be?

```
1 T = (1, 2, 3)
2 print(T * 2)
```

Ungraded 0.0

Question 1 Reset Later**Select an option** Clear Response

What will the output of the following code be?

```
1. def my_func(*args):
2.     return(len(args))
3.
4. print(my_func(1,2,3,4,5))
5. print(my_func(1,2,3))
```

 12

21

 150

210

 10

Error: Invalid number of arguments passed

Dots



Write a query to get 50% out of the following salary replacement
- 12/1/2008, last_name, department
100% salary + 50% (100% of salary) at lead_salary
not replaced

last_name	salary	dept_id
Sutherland	54000	45
Yates	60000	45
Erickson	42000	45
Parker	57500	30
Gates	65000	30

dept_id	last_name	salary	lead_salary
45	Erickson	42000	NULL
45	Sutherland	54000	42000
30	Parker	57500	54000
30	Gates	65000	57500
45	Yates	80000	65000

dept_id	last_name	salary	lead_salary
45	Yates	80000	65000
30	Gates	65000	57500
30	Parker	57500	54000
45	Sutherland	54000	42000
45	Erickson	42000	NULL

? output of the following program be?

```
, 2, 3, 4]
def cube_func(n):
    n**3
list(map(cube_func, [1]))
```

[1, 8, 27, 64]

[1, 2, 3, 4, 1, 2, 3, 4, 1, 2, 3, 4]

[[1, 2, 3, 4], [1, 2, 3, 4], [1, 2, 3,

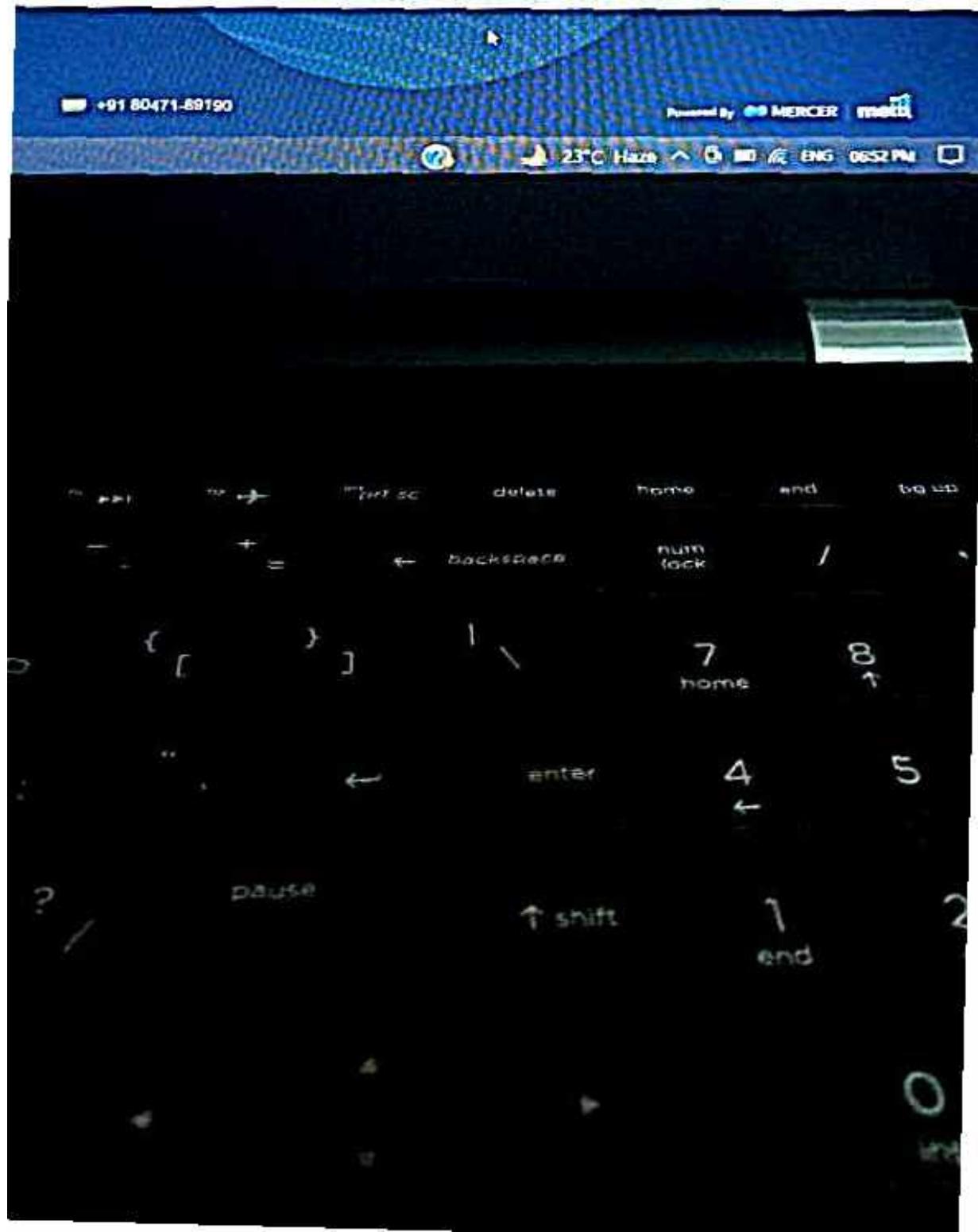
Error



Looks like you are little early here!

This test will start on Feb 06, 2022 at 07:00 PM (Asia/Kolkata).

You will be required to provide some information and present your ID card to verification purpose during the authorisation process.



Attempted (0)

Question 1

 Revise later Select an option Clear Response

What will be the output of the following code?

```
list = ['Hello', 'Hi', 'Hello', 'Hello']  
list.join('Hello')
```

 This is a python test.
 This is a python test. This is a syntax test. None of the above

Question 2

What is the output of the following program? [2]

```
# L = [1, 2, 3, 4]
# def cube_func(x):
#     return x**3
# print(cube_func(cube_func(1)))
```

 Result Given Select an answer

(A) 64

(B) 1, 2, 3, 4

(C) 1, 2, 3, 4, 5, 6

(D) None of the above

 Clear Response

Question 2 Show list

Select an option

 Clear response

What will the output of the following program be?

```
1 L = [1, 2, 3, 4]
2 def print_nums(x):
3     return x**2
4 print(list(map(lambda x: x**2, L)))
```

 (1) 1, 2, 3, 4 (2) 1, 2, 3, 4, 1, 2, 3, 4, 5, 6, 7, 8, 9 (3) 2, 3, 4, 1, 2, 3, 4, 5, 6, 7, 8

Question 3

None of the following code snippets in Python 3.7 give the output `[1, 4, 9, 16, 25]`

 Recompile Select an option Clear Response `list(range(1, 6))` `list([1, 4, 9, 16, 25])` `list(range(1, 6))` `list([1, 4, 9, 16, 25])` None of the above Chat Now

Question 3

What will the following code snippet NOT give me output as [1, 3, 5, 7, 9]?

 Print list Selection option One Response

def printList():
 return [1, 2, 3]

print([*printList()])

print(*printList())

for i in range(1, 10):
 print(i)

None of the above

none of the above

Question 3

 Revisit Later

Select an option

What will be the output of the following:

```
1 a = True
2 b = False
3 c = True
4
5 if not a or b:
6     print "a"
7 elif not a or not b and c:
8     print "b"
9 elif not a or b or not b and a:
10    print "c"
11 else:
12    print "d"
```

 a b c d

Question 1 Print List Select an option Other Response

Suppose you have the following two pandas series:

You want to get the elements in S_1 that aren't present in S_2 .

$S_1 = pd.Series([1, 2, 3, 4, 5, 6, 7, 8, 9])$

$S_2 = pd.Series([1, 2, 3, 4, 5, 6, 10])$

Which of the following command will achieve that?

 $S_1 \setminus S_2$ $S_1 \setminus S_2$ $S_2 \setminus S_1$ $S_2 \setminus S_1$ 

Progress: 0%

Question 2 Review Later

Select an option

 One Response

Suppose you have two separate data frames, df1 and df2 as given below.

These two dataframes are combined in such a way that the resultant dataframe looks as follows:

Name	Age	Gender	Blood Group	Height
S. Vinay	22	M	B+	172
R. Rishi	26	M	O-	165
Z. Ravneet	22	F	A-	168
T. Veena	30	M	B-	160

These two dataframes are combined in such a way that the resultant dataframe looks as follows:

Name	Age	Gender	Blood Group	Height
S. Vinay	24	M	A-	172
R. Rishi	29	M	O-	165
Z. Ravneet	22	F	A-	168
T. Veena	31	M	B-	160

Which of the following command was used to do this?

 pd.append(df2, df1, sort = True) df.append(df2, df1, sort = True) pd.concat(df1, df2, axis = 0) pd.concat(df1, df2, axis = 1)ENG - English (India)
English (India) keyboardENG - English (United States)
US Keyboard Chat Now

Quesiton 2

 Mark Later Select an option Clear Response

What will be the output of the following:

- 1. `a = True`
- 2. `b = False`
- 3. `c = True`
- 4. `If not a or b:
 print('a')`
- 5. `If (not a or not b) and c:
 print('a')`
- 6. `If (not a or b) or not b and c:
 print('a')`
- 7. `None`
- 8. `print('a')`



A yellow hand-drawn heart shape is drawn over the first radio button in the list.

1
2
3
4
5
6
7
8

Question 5 Revisit Later

Which of the following lines of code will NOT throw an error for the given two lists?

```
List1 = [1, 2, 3]  
List2 = [4, 5, 6]
```

Select an option List1 * List2 import numpy as np
np.array(List1)+np.array(List2) List1-List2 import pandas as pd
pd.asarray(L1)+pd.asarray(L2)

Question 4 Run this code Select all options Clear Report

Consider the dataframe "I" provided below. It has the details of highest score made by India's batsmen in ODIs matches.

	Name	Highest_Score	Venue	Opponent
0	M.S.Dhoni	181	Jaypur	Sri Lanka
1	Sachin Tendulkar	206	Chennai	South Africa
2	V.Selvagganesh	215	Moore	West Indies
3	V.Kohli	183	Dhaka	Pakistan
4	Rohit Sharma	264	Kolkata	Srilanka

 All of the above (Name, Highest_Score, Venue, Opponent) All of the above (P 1, 2, 3)
Select all options None of the above

Which of the code snippets will return the row containing the data given below?
(P.Made 100 Runs in Test)

Question 4

由 Squid 完成

Scilink.org

43. Open Systems

Consider the statement (P) provided below. A has the status of right if some track by road or water is open to traffic.

	Name	Highest_Score	Venue	Opponent
0	M.S.Dhoni	183	Jaipur	Sri Lanka
1	Sachin Tendulkar	200	Gwalior	South Africa
2	V.Schwarc	219	Indore	West Indies
3	V.Kohli	183	Dhaka	Pakistan
4	Rohit Sharma	264	Kolkata	Sri Lanka

which of the code snippets will fetch the row containing the data given below?

第二回 喜冤家

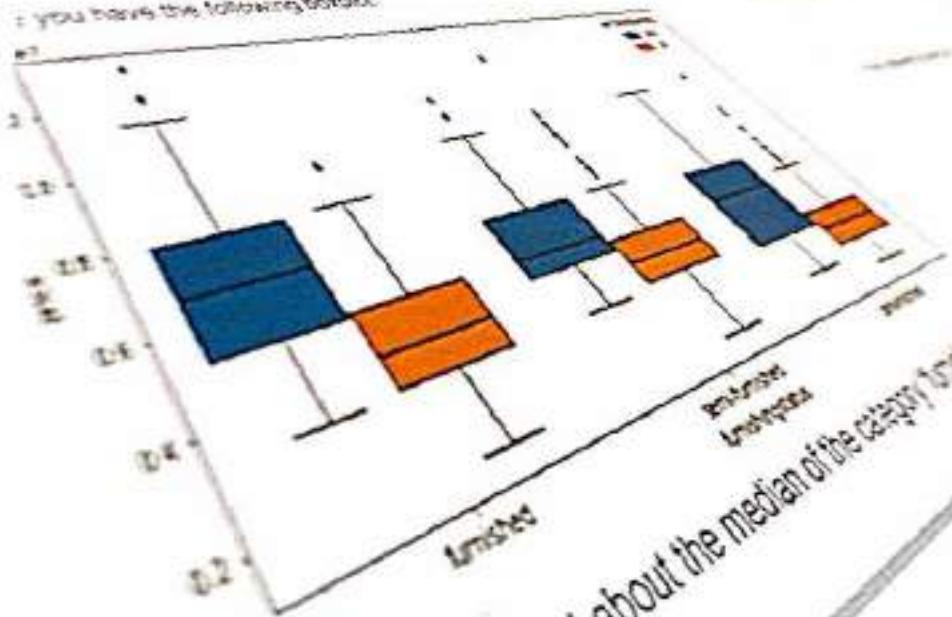


Miyazaki et al. / 2011-2012 303

Want help? Contact us: +1(800)254-8331 | support@wix.com

Chat Now 

You have the following boxplots.

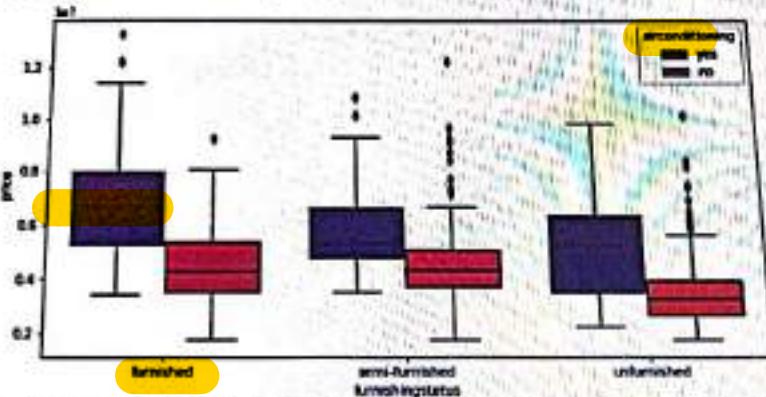


What can be said about the median of the category 'Pilsner'?
Conditioning?

Continuous variable

Question 7

Suppose you have the following boxplot:



Revise Later

Select an option

- It lies between 0.6 and 0.2.
- It lies between 0.6 and 0.4.
- It lies between 0.4 and 0.2.
- It is exactly equal to 0.5.

What can be said about the median of the category 'furnished with air-conditioning'?

tests.metli.com is sharing your screen.

Question 3 Revisit Later

Select an option

Consider the following table and answer the questions.

Given below is the data of school boys and girls of an Indian state who reported the number of times they play games, i.e. whether they play every day, never, once a month or once a week.

	Everyday	Never	Once a month	Once a week	Total
Boy	3474	154	150	780	4558
Girl	2886	175	200	1046	4307

Approx. what percentage of girls play at least once a week over the total no. of girls?

 3% 91% 35% 44%

Question 1

 Partial Score

Select an option

 Clear Response

You are an analyst working for Netflix and trying to figure out the kind of factors that affect the 'Viewer Rating' for a recent crop of 21 Netflix original movies and you prepared the following correlation matrix:

	"Budget (M\$)"	"Genre_Language (M\$)"	"Movie_Duration (min)"	"Viewer Rating"
"Budget (M\$)"	1			
"Genre_Language (M\$)"	0.45	1		
"Movie_Duration (min)"	0.1	0.23	1	
"Viewer Rating"	0.51	0.55	-0.13	1

Now answer the following questions.

Which of the following attributes mainly effect the 'Viewer Rating'?

 Cross rating Movie Length Both a and b Budget only



Question 3 Revise Later

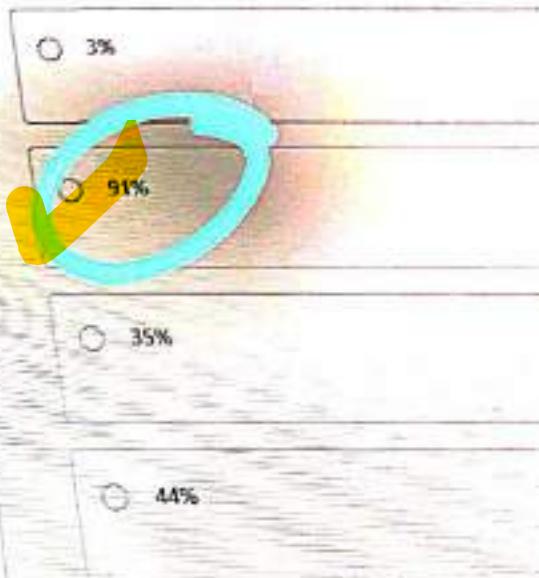
Select an option

Consider the following table and answer the questions.

Given below is the data of school boys and girls of an Indian state who reported the number of times they play games, i.e. whether they play every day, never, once a month or once a week.

	Everyday	Never	Once a month	Once a week	Total
Boy	3474	154	150	780	4558
Girl	2886	175	200	1046	4307

Approx. what percentage of girls play at least once a week over the total no. of girls?



Attempted 0/8

Question 2

 Prev Question Select an option Our Response

You are an analyst working for Netflix and trying to figure out the kind of factors that affect the "Viewer Rating" for a number of 25000 movies. You prepared the following correlation matrix.

	"Budget (EMD)"	"Gross Earnings (M\$)"	"Movie Length (min)"	"Viewer Rating"
"Budget (EMD)"	1			
"Gross Earnings (M\$)"	0.45	1		
"Movie Length (min)"	-0.1	0.25	1	
"Viewer Rating"	0.08	0.05	0.22	1

Consider the following two statements:-

- (i) "Movie Length" and "Gross Earnings" are inversely related since they have a negative correlation value.
- (ii) "Movie Length" and "Budget" are inversely related since they have a very low correlation value.

Given the above statements, choose the most appropriate option.

 Statement 1 is correct. Statement 2 is false. Statement 1 is false. Statement 2 is correct. Both statements are correct. Both statements are incorrect.

Question 4

 Previous Selected answer Clear Response

Which of the following visualizations is suited for univariate analysis?

 Both A and BENG English (US)
English (United Kingdom)ENG English (United States)
US keyboard

Attempted: 0/9

Revisit Later

Select an option

nts from a particular college have applied for a job at a particular
y the probability that a student from that college gets that job is 0.25.
ility that more than 2 people from that college are selected for the job.

39%

61%

55%

45%

Question 2

How to Attempt?

Given a list of numbers, find out if 3 or 2
divisible by either 4 or 9. Output ~~numbers~~

Examples:



Input 1:

[3, 8, 12, 15, 18, 9, 4, 27, 36, 40]

Output 1:

[8, 12, 18, 9, 4, 27, 36, 40]

Note: The order in the input list is maintained

Expected Output

Expected Output:
A Boolean value (True or False) based on whether or not.

Examples:

Input 1:

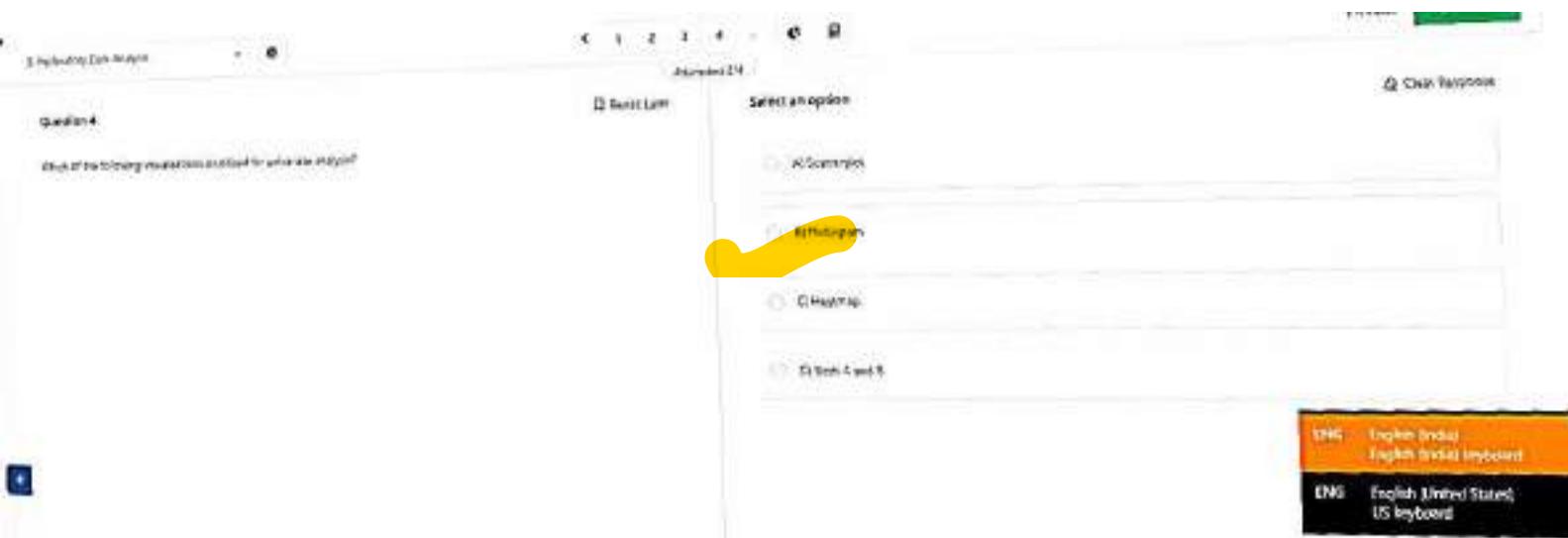
Output 1:

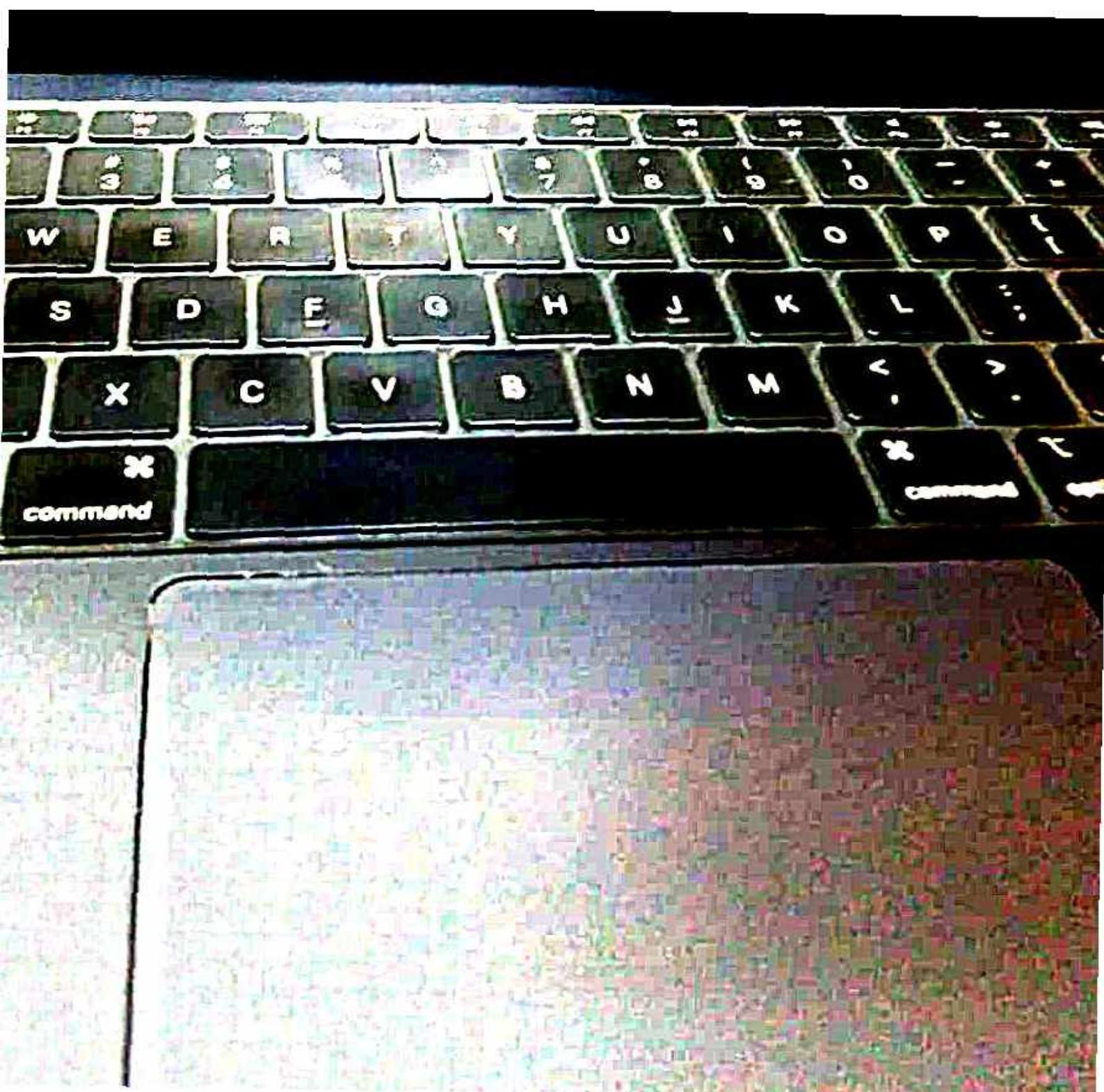
Explanation:
0 0

Explanation:

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & -1 \\ 1 & 0 & -1 & -1 & 0 \\ \textcolor{red}{0} & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

As we have a dear past, you can often meet
encountering -1, you can often meet



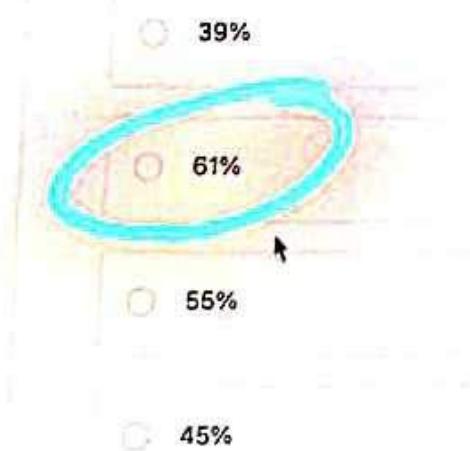


Attempted: 0/9

Revisit Later

Select an option

nts from a particular college have applied for a job at a particular
y the probability that a student from that college gets that job is 0.25.
ility that more than 2 people from that college are selected for the job.



Attempted: 1/7

Question 2

 Revisit Later

Select an option

Consider the following tables containing data about the marks obtained by students in three courses, namely, Mathematics, Science and Economics and answer the questions that follow.

The names of the tables are Marks, Students and Teachers respectively.

Marks

Student_ID	Marks	Course
32	99	Mathematics
22	91	Mathematics
12	99	Mathematics
17	100	Mathematics
3	88	Mathematics
32	97	Science
22	57	Science
12	91	Science
17	91	Science
3	87	Science

```
select avg(Marks)
from Marks
where Course = 'Economics'
```

```
select avg(marks)
from Marks
where Course != 'Economics';
```

```
select avg(marks)
from marks
where course = economics;
```

```
select avg(marks)
from marks
where course or 'economics';
```

#	#	Course
12	91	Science
17	91	Science
3	87	Science
32	90	Economics
22	87	Economics
12	71	Economics
17	80	Economics
3	89	Economics

Students

Student_Name	Student_ID	Gender
Ajay Gupta	12	Male
Neeraja Ingle	22	Female
Mohima Prasad	17	Male
Harshit Kati	32	Male
Akash Giriya	3	Male

Teachers

Name	ID	Age	Course_Taught
Kishan Jain	19	24	Mathematics
Reetesh Chandra	18	25	Science
Abd Surana	7	27	Economics

Which of the following queries will display the average marks obtained in the Economics course?

a) select avg(Marks) from Marks where Course = 'Economics'

b) select avg(Marks) from Marks where course = 'Economics'

c) select avg(Marks) from Marks where course = 'Economics'

Out line

Attempted 6/8

Question 7 Review Later Select an option: Clear Response

Consider the following hypothesis statement:

 $H_0: \mu = 150$ $H_1: \mu < 150$

The sample size is 265. Sample standard deviation is 15 and the sample size is 26. Calculate the p-value and make a decision at the level of significance of 5%. Use the Z-table given below.

-z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5048	.5080	.5110	.5140	.5169	.5199	.5229	.5259	.5289
0.1	.5398	.5438	.5470	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6629	.6664	.6700	.6736	.6772	.6809	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7325	.7357	.7391	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8105	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9293	.9306	.9319
1.5	.9332	.9346	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9776	.9781	.9788	.9793	.9798	.9803	.9808	.9812	.9817

 15.2% Fail to reject the null hypothesis 2.5%, Fail to reject the null hypothesis 2.7%, reject the null hypothesis 4.1%, reject the null hypothesisENGLISH (United Kingdom)
English (United Kingdom)ENGLISH (United States)
US keyboard

Question 9 Review Later See Response

The daily average time spent by customers on Netflix in the Hollywood movies section is 15 hours. The India Head of Netflix wants to verify whether its has similar Indian customers. Then there plans to reduce the daily average duration spent to let an customer in the Hollywood movies section to be used as a sample to determine whether it is significantly different from 15 hours.

The following hypothesis statements are used:

H₀: $\mu = 15$

H_A: $\mu \neq 15$

Now consider the following statements regarding the kind of errors that can happen in the above hypothesis test:

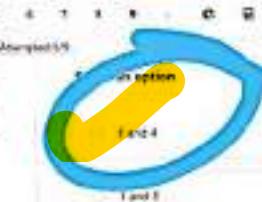
Statement 1: A Type-I error happens when the India Head of Netflix concludes that the average duration spent by Indian customers in the Hollywood movies section is less than 15 hours when it actually is.

Statement 2: A Type-II error happens when the India Head of Netflix concludes that the average duration spent by Indian customers in the Hollywood movies section is 15 hours when it actually is not.

Statement 3: A Type-I error happens when the India Head of Netflix concludes that the average duration spent by Indian customers in the Hollywood movies section is 15 hours when it actually is.

Statement 4: A Type-II error happens when the India Head of Netflix concludes that the average duration spent by Indian customers in the Hollywood movies section is 15 hours when it actually is not.

Which of the above statements are correct?



ENG English (British)
English (India) keyboard
FNG French (France) keyboard
US keyboard

Ask question asking your issue

Next Question

Outline

Question 4

What will be the output of the given SQL query for the following table employees?

- 1. SELECT emp_id, last_name, salary;
- 2. SELECT last_name, MAX(salary) AS max_salary, MIN(salary) AS min_salary;
- 3. SELECT employees;

last_name	salary	dept_id
Sutherland	54000	45
Yates	60000	45
Erickson	42000	45
Parker	57500	30
Gates	65000	30

D. Result List

Answered 100%

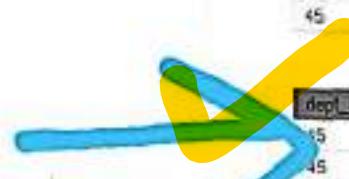
dept_id	last_name	salary	lead_salary
45	Erickson	42000	NULL
45	Sutherland	54000	42000
30	Parker	57500	54000
30	Gates	65000	57500
45	Yates	60000	65000

ENG English-English
English-Arabic keyboard
ENG English-United States
US keyboard

dept_id	last_name	salary	lead_salary
45	Yates	60000	65000
30	Gates	65000	57500
30	Parker	57500	54000
45	Sutherland	54000	42000
45	Erickson	42000	NULL

dept_id	last_name	salary	lead_salary
45	Yates	60000	NULL
30	Gates	65000	50000
30	Parker	57500	65000
45	Sutherland	54000	57500
45	Erickson	42000	54000

dept_id	last_name	salary	lead_salary
45	Erickson	42000	54000
45	Sutherland	54000	57500



A SQL Query

		Mark_Marks
12	91	Science
17	91	Science
3	87	Science
32	90	Economics
22	87	Economics
12	71	Economics
17	90	Economics
3	89	Economics

Students

Student_Name	Student_ID	Gender
Ajay Gupta	12	Male
Neeta Jingle	22	Female
Mahima Prasad	17	Male
Harshit Kati	32	Male
Akshay Ghoda	3	Male

Teachers

Name	Id	Age	Courses_Taught
Kshitij Jain	19	24	Mathematics
Rakesh Gandra	18	35	Science
Aditi Suman	7	27	Economics

Select all the following queries and find the names of all students who have scored 99 marks in Mathematics (More than one option may be correct).

1. select student_name from students
where student_id = student_id
and marks > 99 and course = 'Mathematics'

2. select student_name
from students
where student_id = student_id
and marks > 99 and course = 'Mathematics'

3. select student_name
from students
where student_id = student_id
and marks > 99 and course = 'Mathematics'

only A

1. select student_name from students
where student_id = student_id
and marks > 99 and course = 'Mathematics'

Chat Now



▷ ↴

```
input1=[3,8,12,15,18,9,4,27,36,40]
output1=[8,12,18,9,4,27,36,42]
print(list(filter(lambda x: (x%4==0) or (x%9==0), input1)))
#[8, 12, 18, 9, 4, 27, 36, 40]
```

[26]

✓ 0.9s

... [8, 12, 18, 9, 4, 27, 36, 40]

L.WJ.1.1

Attempted 21

Q. Order Date Response

HR100,00

```

1 select count(Order_ID) as M, Order_Date
2 from Orders
3 group by Order_Date
4 order by M, Order_Date
5 limit 10;
    
```

ANSWER

- Calculate count of all the orders.
- With the alias of **M** for count of orders.
- Group the results by **Order_Date**.
- Order items based on **Order_Date** in **descending order**.
- Limit to 10 results.

OUTPUT COLUMN

as **Order_Dates**

Here is an image showing how a sample output would look like:

OC	ORDER_DATE
1	2015-11-03
2	2019-03-18

~_ CONSOLE OUTPUT

STANDARD ERROR/WARNING

Traceback (most recent call last):

File "UserMainCode.py", line 17, in treasure
for i in input1:

TypeError: 'int' object is not iterable

— CONSOLE OUTPUT

STANDARD ERROR/WARNING

Traceback (most recent call last):

File "UserMain.py", line 1:

```
print(list(filter(lambda x: x <= 0, <input1>)))  
          ^
```

SyntaxError: invalid syntax

Sample 2 .

2 Previous

Attempted: 2/2

ON3

Compiler: Python 3.6

:

```
# Read only region start
class userMainCode(object):
    @classmethod
    def divisibility_list(cls, input1):
        ...
        input1 : List[Integer]

    Expected return type : List[Integer]
    ...

# Read only region end
# Write code here

    return(list(filter(lambda x: (x%4==0) or (x%9==0), input1)))
```

Examples:

Input 1:

4, 5, [0,0,0,0,0,1,0,-1,0,-1,0,0,-1,-1,-1,0,0,2,0,0]

Output 1:

True

Explanation:

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 & -1 \\ 0 & 0 & -1 & -1 & -1 \\ 0 & 0 & 2 & 0 & 0 \end{bmatrix}$$

As we have a clear path from 1 to 2 without encountering -1,
you can get to the treasure.

Input 1:

4, 5, [0,0,0,0,0,1,0,-1,0,-1,0,0,-1,-1,-1,0,-1,2,0,0]

Output 1:

False

to Attempt?

Imagine you are playing a game in which you need to find a treasure. You are given a matrix with your position denoted by '1' and the treasure's position denoted by '2'. All other entries in the matrix will be either 0 (safe tile) or -1 (unsafe tile). Can you get the treasure or not?

You can travel up, down, left or right only (no diagonals).

Input will be given as:

the number of rows, number of columns, [all the matrix entries in the form of a list of lists]

Expected Output:

A Boolean value (True or False) based on whether you get the treasure or not.



Examples:

Input 1:

4, 5, [0, 0, 0, 0, 0, 1, 0, -1, 0, 0, -1, -1, -1, 0, 0, 2, 0, 0]

ans.net



```
# Write code here
return(list(filter(lambda x:(x%4==0) or (x%9==0), input1)))
```

I

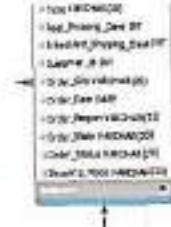
Revisit Later

are divisible

Attempted 1/2

```
6
7     input1 : List[Integer]
8
9     Expected return type: List[Integer]
10    ...
11     # Read only region end
12     result=[]
13     for i in input1:
14         if(i%4==0 or i%9==0):
15             result.append(1)
16     return result
17
18
19
```

 Use Custom Input



QUERY

- Calculate count of all the orders.
- Where OrderDate is March 2010.
- With - Give the class of the transaction.
- Group the results by Type.
- Order them by count ascending order.

OUTPUT COLUMNS

as Type

Here is an image showing how a sample output would look like:

OC	TYPE
45	CASH
89	PAYOUT
108	TRANSFER
151	DEBIT

Note that the resulting results automatically converts the name of the columns to upper case.

6. Python Coding

How to Answer?

Imagine you are playing a game in which you need to find a treasure. You are given a matrix with your position denoted by "0" and the treasure's position denoted by "2". All other entries in the matrix will be either 0 (safe tile) or -1 (unsafe tile). Can you get the treasure or not?

You can travel up, down, left or right only (no diagonals).

Input will be given as:
the number of rows, number of columns, [all the matrix entries in the form of a list of lists]

Expected Output:

A Boolean value (True or False) based on whether you get the treasure or not.

Examples:

Input 1:
4,3,[0,0,0,0,1,-1,0,-1,0,0,-1,-1,0,0,2,0,0]

Output 1:

True

Explanation:

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 & -1 \\ 0 & 0 & -1 & -1 & -1 \\ 0 & 0 & 2 & 0 & 0 \end{bmatrix}$$

As we have a clear path from 1 to 2 without encountering -1, we can get to the treasure.

Input 2:
4,5,[0,0,0,0,0,1,0,-1,0,1,-1,0,-1,0,-1,0,0,0,0]

Output 1:

False

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 & -1 \\ 0 & 0 & -1 & -1 & -1 \\ 0 & -1 & 2 & 0 & 0 \end{bmatrix}$$

The screenshot shows a Python code editor with the following code:

```

# Read only region start
class Solution(object):
    @lru_cache(maxsize=None)
    def hasPath(self, matrix, start, input1):
        ...
        input1 : int
        input2 : int
        input3 : List[int]
    ...
    expected return type : bool
    ...
    # Read only region end
    # Write code here
    pass

```

Below the code editor is a terminal window with the following output:

```

$ python3 main.py
[[0,0,0,0,0,1,0,-1,0,1,-1,0,-1,0,-1,0,0,0,0], 4,5]
True

```

At the bottom of the interface, there are buttons for "Compile and Test" and "Chat Now".

QUESTION

You're given a database and the columns in the fact table are shown below:

Order

- Order_ID
- Order_Value
- Order_Status
- Customer_Identifier
- Order_Date
- Order_Year
- Order_Month
- Order_Day
- Order_Hour
- Order_Minute
- Order_Second
- Order_Millisecond
- Order_Microsecond
- Order_Nanosecond

ANSWER

MySQL

Q1: Enter your Response

MySQL

QUERY

- Calculate count of all the orders.
- With - Use the alias of fact for count of orders.
- Group the results by Order_Date
- Order them by id & Order_Date in ascending order
- Limit to 10 results

OUTPUT/COLUMNS

as Order_Count

Here's an image showing how a sample output would look like:

OC	ORDER_DATE
1	2018-11-01
1	2019-03-18
1	2020-12-10
2	2018-11-11
2	2018-11-29

5:00: M00

6 1 2 3 4 5 6 7 8 9 10

Previous Next

Question 6 Review later Other Response

Select which of the statements is true regarding user-defined functions and stored procedures.



A user-defined function cannot call another user-defined function.

A stored procedure must return a value.

A user-defined function accepts both the input and output parameters.

ENG English (India)
English (India) keyboardHNG French (United States)
US keyboard

Question 8**II. Solution****Selection option****Q. Give Response**

Identify which of the statements is true regarding user-defined functions and stored procedures.

A user-defined function cannot call a stored procedure.

A stored procedure cannot call a user-defined function.

A stored procedure must return a value.

A user-defined function supports both the input and output parameters.

ENG English (India)
English (India) keyboard

DHL English (United States)
US keyboard

Question 5 Revisit Later

What will be the output of the given SQL query for the following table 'employees'?

```
1 SELECT dept_id, last_name, salary,  
2 LEAD (salary,1) OVER (ORDER BY salary) AS lead_salary  
3 FROM employees;
```

last_name	salary	dept_id
Sutherland	54000	45
Yates	80000	45
Erickson	42000	45
Parker	57500	30
Gates	65000	30

ID	Age	Subject
12	91	Science
17	91	Science
3	97	Science
32	90	Economics
22	67	Economics
52	71	Economics
17	99	Economics
3	89	Economics

Students

Student_Name	Student_ID	Gender
Ajay Gupta	12	Male
Neeraja Ingle	22	Female
Mohima Prasad	17	Male
Harshit Kant	32	Male
Ananya Ghodia	3	Male

Teachers

Name	W	Age	Course_Taught
Ranbir Jain	19	24	Mathematics
Reetesh Chandra	16	35	Science
Aditi Savama	7	27	Economics

Which of the following queries will display the average marks obtained in the economics course?

www.w3schools.com/sql/ [Registration](#) [Help](#)

Chat Now

Question 4

A CASE statement in SQL is equivalent to which of the following?

 Review Later Select an option Clear Response

A way to use CASE based logic in SQL.

 A way to use IF-THEN-ELSE logic in SQL

A way to define CASE while creating tables

A way to use FUNCTIONS in SQL

Question 8

Previous Letter

Select answer options

The Screen Sports Industries has been producing running shoes with an average weight over a long period of time. It is thought that the average weight of a running shoe produced by the company is equal to 1.21 kg. A rear analyst is asked who challenges the status quo. The analyst takes a sample of 20 balls with a sample mean of 1.24 kg and a standard deviation of 0.3. Calculate the p-value for the test at the significance level of 5%. You can use the table given below.

	20	21	22	23	24	25	26	27	28	29
6.0	500	540	580	610	640	590	520	570	510	550
6.1	520	560	540	510	530	570	500	550	510	530
6.2	570	510	540	560	530	580	540	510	560	520
6.3	670	617	655	633	668	648	640	660	633	655
6.4	654	692	684	633	678	611	688	664	667	654
6.5	695	660	695	619	664	700	631	617	676	654
6.6	727	731	724	737	742	744	746	717	759	737
6.7	760	741	760	784	778	784	776	763	782	750
6.8	766	710	767	755	762	801	767	716	761	766
6.9	828	811	828	834	826	825	826	835	826	828
7.0	813	828	841	841	825	820	834	827	835	841
7.1	840	865	868	858	875	840	870	870	860	840
7.2	849	860	838	867	860	864	860	860	861	849
7.3	860	869	862	869	853	841	867	842	867	860
7.4	872	857	821	856	879	832	856	859	857	872
7.5	873	845	877	882	884	846	848	845	884	873
7.6	882	862	844	846	855	812	852	852	854	882
7.7	894	864	873	853	881	869	868	865	862	894
7.8	895	869	864	867	867	868	860	869	876	895
7.9	872	859	856	858	874	876	875	871	877	872
8.0	877	879	873	870	879	881	890	861	887	877

 10.12% fail to reject the null hypothesis 5.59% fail to reject the null hypothesis 5.34% reject the null hypothesis 10.12% reject the null hypothesis

Chat live

www.meritnation.com is sharing your answers.

Next Question

SQLMCQ

0

4 1 2 3 4 5 6 7 8 9 10

Previous Next

Id	StID	Subject
12	91	Science
17	81	Science
3	87	Science
32	90	Economics
22	87	Economics
12	71	Economics
17	90	Economics
3	69	Economics

Students

Student_Name	Student_ID	Gender
Ajay Gupta	12	Male
Neeraja Ingle	22	Female
Mohima Prasad	17	Male
Harshit Kant	32	Male
Akshay Ghoda	3	Male

Teachers

Name	Id	Age	Course_Taught
Kabita Jain	18	24	Mathematics
Reetash Chandra	16	36	Science
Adil Suvana	7	27	Economics

Which of the following queries will fetch the names of all students who have scored 99 marks in Mathematics? More than one option may be correct.

Answered 37

Choose the best option

```
select Student_Name
from Students
inner join Marks
on Student_ID = student_ID
where Marks > 99 and Course = 'Mathematics'
```

```
select Student_Name
from Marks
inner join Students
on Student_ID = student_ID
where Marks > 99 and Course = 'Mathematics'
```

```
select Student_Name
from Students
inner join Marks
on student_ID = student_ID
where Marks > 99 and Course = 'Mathematics'
```

```
select Student_Name
from Students
inner join Marks
on student_ID = student_ID
where Marks > 99 and Course = 'Mathematics'
```

@ GM Support

@ GM Note

Question 8

 Revisit Later

Select an option

The Screen Sports Industries has been producing cricket bats with its stable supply chain for a long period of time. It is thought that the average weight of a cricket bat produced by the company is equal to 1.23 kg. A new analyst is hired who challenges the status quo. The analyst takes a sample of 49 bats with a sample mean of 1.3 kg and a standard deviation of 0.3, respectively. Calculate the p-value for the test at the significance level of 5%. You can use the z-table given below.

<i>Z</i>	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09	.10
-1.9	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359	
-1.8	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753	
-1.7	.5711	.5752	.5791	.5831	.5871	.5910	.5949	.5987	.6026	.6064	.6103
-1.6	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517	
-1.5	.6524	.6561	.6628	.6644	.6700	.6736	.6772	.6808	.6844	.6879	
-1.4	.6915	.6953	.6995	.7039	.7074	.7088	.7123	.7157	.7190	.7224	
-1.3	.7272	.7311	.7341	.7371	.7395	.7422	.7454	.7486	.7517	.7549	
-1.2	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852	
-1.1	.7881	.7912	.7939	.7967	.7991	.8013	.8031	.8051	.8073	.8106	.8133
-1.0	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389	
-0.9	.8411	.8428	.8441	.8458	.8473	.8488	.8504	.8519	.8537	.8559	.8621
-0.8	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830	
-0.7	.8845	.8869	.8893	.8917	.8935	.8954	.8972	.8990	.9007	.9015	
-0.6	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177	
-0.5	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319	
-0.4	.9272	.9295	.9317	.9337	.9352	.9374	.9406	.9418	.9429	.9441	
-0.3	.9343	.9361	.9374	.9384	.9405	.9425	.9451	.9475	.9505	.9545	
-0.2	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633	
-0.1	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706	
0.0	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767	
0.1	.9772	.9772	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817	

10.32%; fail to reject the null hypothesis

5.16%; fail to reject the null hypothesis

5.16%; reject the null hypothesis

10.32%; reject the null hypothesis

Consider the following tables containing data about the marks obtained by students in three courses, namely, Mathematics, Science and Economics and answer the questions that follow.

The names of the tables are Marks, Students and Teachers respectively.

Marks

Student_ID	Marks	Course
32	99	Mathematics
22	91	Mathematics
12	99	Mathematics
17	100	Mathematics
3	88	Mathematics
32	97	Science
22	57	Science
12	91	Science
17	91	Science
3	87	Science
..

Choose the best option

select Student_ID, Course, Student_name from Marks inner join Students using(Student_ID) where Course = 'Mathematics' and student_name = 'John';
 select max(Marks) from Marks;

select Student_ID, course, student_name from marks inner join students using(student_id) where course = 'mathematics' and student_name = 'John';
 select marks from marks;

select m.Student_ID, Course, Student_name from Marks m inner join Students s on m.Student_ID = s.Student_ID where Course = 'Mathematics' and student_name = 'John';
 select max(Marks) from Marks;

Question 2

 Revise later

Consider the following tables containing data about the marks obtained by students in three courses, namely, Mathematics, Science and Economics and answer the questions that follow.

The names of the tables are Marks, Students and Teachers respectively.

Marks

Student_ID	Marks	Course
32	99	Mathematics
22	91	Mathematics
12	99	Mathematics
17	100	Mathematics
3	88	Mathematics
32	97	Science
22	57	Science

select avg(Marks)
from Marks
where Course = 'Economics'

select avg(marks)
from Marks
where Course != 'Economics';

select avg(marks)
from marks
where course = economics;

select avg(marks)
from marks
where course of "economics";

!! tests.mettl.com is sharing your screen.

 Stop sharing Hide

Question 4 Revisit Later Select an option

The dean of a college wants to know the average time spent (in hours) by the students of his college in the library. He takes a sample of 225 students visiting the library and calculates the mean time and standard deviation, which comes out to be 6 and 1.5, respectively. Find the interval in which the average time spent in the library by the entire population might lie with a confidence level of 90%. You can use the Z-table given below.

(7.531, 8.110)

(5.5, 7.5)

<i>z</i>	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817



Consider the following tables containing data about the marks obtained by students in three courses, namely, Mathematics, Science and Economics and answer the questions that follow.

Names of the tables are Marks, Students and Teachers respectively.

Marks

student_ID	Marks	Course
32	99	Mathematics
22	91	Mathematics
12	99	Mathematics
17	100	Mathematics
3	88	Mathematics
32	97	Science
22	57	Science
12	91	Science
17	91	Science
3	87	Science
--	--	--

Choose the best option

select Student_ID, Course, Student from Marks inner join Students using(Student_ID) where Course = 'Mathematics' and select max(Marks) from Marks;

select Student_ID, course, student from marks inner join students using(student_id) where course = 'mathematics' and m select marks from marks;

select s.Student_ID, Course, Student from Marks m inner join Students s on m.Student_ID = s.Student_ID where Course = 'Mathematics' and select max(Marks) from Marks;

Question 8

 Revisit Later

Select an answer

The Sarwan Sports Industries has been producing cricket bats with its stable supply chain for a long period of time. It is thought that the average weight of a cricket bat produced by the company is equal to 1.23 kg. A new analyst is hired who challenges the status quo. The analyst takes a sample of 49 bats with a sample mean of 1.3 kg and a standard deviation of 0.3, respectively. Calculate the p-value for the test at the significance level of 5%. You can use the z-table given below.

-	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.3	5000	5040	5080	5120	5160	5199	5239	5279	5319	5359
0.4	5398	5438	5478	5517	5557	5596	5636	5675	5714	5753
0.5	5792	5832	5871	5910	5949	5987	6026	6064	6103	6141
0.6	6179	6217	6255	6293	6331	6368	6406	6443	6480	6517
0.7	6654	6691	6728	6764	6790	6796	6772	6808	6844	6879
0.8	6915	6952	6985	7019	7054	7088	7123	7157	7190	7224
0.9	7252	7281	7314	7342	7370	7404	7434	7466	7517	7549
1.0	7600	7641	7682	7673	7704	7734	7764	7794	7823	7852
1.1	7981	8021	8059	8087	8126	8163	8191	8219	8106	8133
1.2	8309	8356	8212	8238	8264	8289	8315	8340	8365	8389
1.3	8611	8658	8641	8645	8658	8631	8654	8677	8599	8621
1.4	8963	8965	8966	8798	8729	8749	8770	8790	8810	8830
1.5	9049	9099	9098	9167	9125	9144	9162	9180	9197	9015
1.6	9032	9049	9066	9062	9099	9115	9131	9147	9162	9177
1.7	9150	9227	9222	9236	9251	9265	9279	9292	9306	9319
1.8	9322	9345	9357	9370	9382	9394	9406	9418	9429	9441
1.9	9452	9451	9474	9484	9476	9505	9515	9525	9535	9545
2.0	9554	9564	9573	9582	9591	9599	9608	9616	9625	9633
2.1	9641	9649	9656	9664	9671	9678	9686	9693	9699	9706
2.2	9713	9719	9726	9732	9738	9744	9750	9756	9761	9767
2.3	9772	9778	9793	9788	9793	9798	9803	9808	9812	9817

10.32%; fail to reject the null hypothesis

5.16%; fail to reject the null hypothesis

5.16%; reject the null hypothesis

10.32%; reject the null hypothesis

Question 4

 Review Later

Select an option

A CASE statement in SQL is equivalent to which of the following?

A way to use CASE based logic in SQL

A way to use IF-THEN-ELSE logic in SQL

A way to define CASE while creating tables

A way to use FUNCTIONS in SQL



Attempted 0/9

Question 9 Randomized Select an option **Get Response**

The daily average time spent by customers on Netflix in the Hollywood movies section is 1.5 hours. The India Head of Netflix wants to verify whether this holds true for Indian customers. Their team plans to record the daily average duration spent by Indian customers in the Hollywood movies section. Is he using a sample to determine whether it is significantly different from 1.5 hours?

The following hypothesis statements are used:
 $H_0: \mu = 1.5$
 $H_A: \mu \neq 1.5$

Now consider the following statements regarding the types of errors that can happen in the above hypothesis test:

Statement 1: A Type-I error happens when the India Head of Netflix concludes that the average duration spent by Indian customers in the Hollywood movies section is not 1.5 hours when it actually is.

Statement 2: A Type-II error happens when the India Head of Netflix concludes that the average duration spent by Indian customers in the Hollywood movies section is not 1.5 hours when it actually is not.

Statement 3: A Type-I error happens when the India Head of Netflix concludes that the average duration spent by Indian customers in the Hollywood movies section is 1.5 hours when it actually is not.

Statement 4: A Type-II error happens when the India Head of Netflix concludes that the average duration spent by Indian customers in the Hollywood movies section is 1.5 hours when it actually is not.

Which of the above 4 statements are correct?


 1 and 4

 1 and 3

 2 and 3

 2 and 4

ENGLISH English (US)
 English (India) keyboard

DAISY English (United States)
 DAISY (EN)

Question 7 Revisit Later

Consider the following hypothesis statement:

$$H_0: \mu = 250$$

$$H_1: \mu \neq 250$$

The sample mean is 265. Sample standard deviation is 65 and the sample size is 36.

Calculate its p-value and make a decision at the level of significance of 5%. Use the Z-table given below

<i>z</i>	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608			
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686			

Attempted: 4/9

Question 5 Revisit Later**Select an option**

Which of the following sample sizes would result in the largest value of the standard error?
(Assuming that the standard deviation = 2.3)

 49

100

36

225

4 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15

Attempt 2/9

Question 3

 Recall Later

Select an option

 Clear Response

For a particular experiment, the following table was prepared. Here, the first column contains values of the random variable X and the second column contains the associated probabilities. However, one particular value and its associated probability are missing from the data. If the Expected Value of the random variable X came out to be 20.7, calculate both the missing value and its associated probability.

X (values that random variable X can take)	Probability
1	0.01
2	0.07
3	0.17
4	
5	0.17

 $X=0, P=0.7$ $X=2, P=0.7$ $X=4, P=0.7$ $X=6, P=0.7$

Check now

T: testmetton is sharing your screen.

Question 3

Results

For a random variable that is normally distributed the mean comes out to be 6 and the standard deviation comes out to be 1. In any given experiment, what would be the probability that the value of this random variable lies between 4 and 8? You can use the Z-tables given below.

<i>Z</i>	.09	.08	.07	.06	.05	.04	.03	.02	.01	.005	.001	.0005
-3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003
-3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0004	.0004	.0004
-3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005	.0005
-3.1	.0010	.0009	.0009	.0008	.0008	.0008	.0008	.0008	.0007	.0007	.0007	.0007
-3.0	.0013	.0013	.0013	.0012	.0012	.0012	.0011	.0011	.0011	.0011	.0010	.0010
-2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0015	.0015	.0014	.0014
-2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0021	.0020	.0019	.0019
-2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0029	.0028	.0027	.0026	.0026
-2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0039	.0038	.0037	.0036	.0036
-2.5	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0050	.0049	.0048	.0047	.0046
-2.4	.0080	.0079	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0066	.0064	.0064
-2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0086	.0084	.0084
-2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110	.0108	.0107
-2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143	.0140	.0138
-2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0186	.0180	.0175	.0173
-1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233	.0227	.0221
-1.8	.0350	.0345	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294	.0289	.0283

	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9685	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767

- 68%
- 99.3%
- 95.4%
- 97.5%

www.123RF.com

Attempted: 1/9

 Revisit Later

Select an option

variable that is normally distributed the mean comes out to be 6 and the standard deviation comes out to be 1. In any given experiment, what would be the probability that the value of this random variable lies between 4 and 8?

68%

You can use the following table:

.01	.02	.03	.04	.05	.06	.07	.08	.09
.001	.0003	.0001	.0001	.0001	.0001	.0001	.0001	.0002
.005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
.017	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
.009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
.13	.0011	.0012	.0012	.0011	.0011	.0011	.0010	.0010
.18	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
.25	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
.34	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
.45	.0044	.0044	.0044	.0040	.0039	.0038	.0037	.0036
.50	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
.60	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
.74	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
.85	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
.91	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
.92	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
.93	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
.94	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294

99.7%

95.4%

97.5%

.02	.03	.04	.05	.06	.07	.08	.09
.9856	.9854	.9851	.9848	.9846	.9843	.9840	.9836
.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
.9783	.9785	.9783	.9781	.9781	.9780	.9782	.9787
.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
.9858	.9871	.9875	.9878	.9881	.9884	.9887	.9890
.9988	.9901	.9904	.9906	.9909	.9911	.9913	.9916
.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
.9982	.9983	.9984	.9984	.9985	.9985	.9985	.9985

+91 80471-89190

Need Help? Contact us: +1 (800) 265-6038

Attempted: 2/9

Question 2

 Visit Later

Select an option

 Clear Response

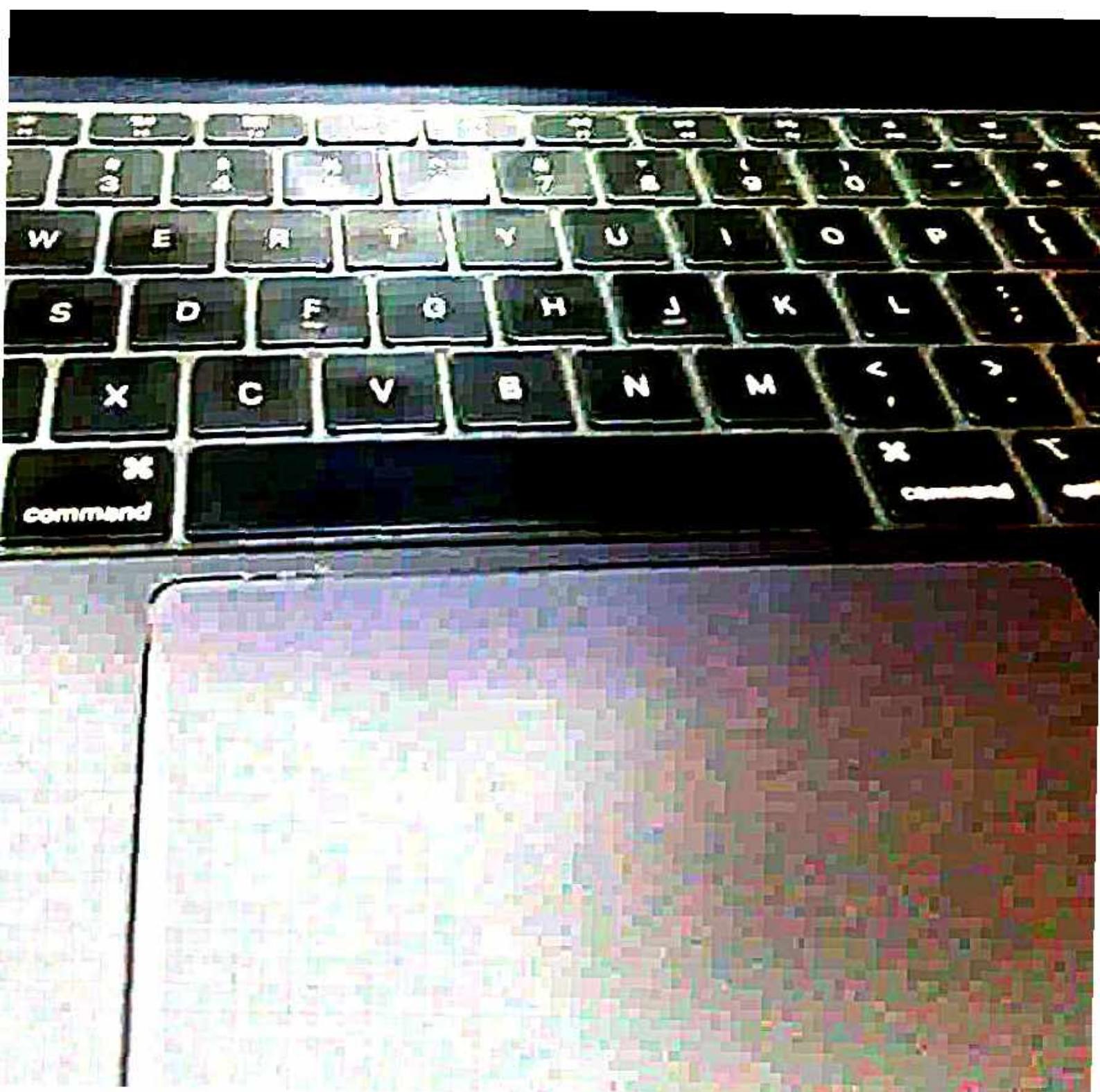
For a random variable that is normally distributed the mean comes out to be -1 and the standard deviation comes out to be 1. In any given experiment, what would be the probability that the value of this random variable lies between 4 and 8? You can use the Z-tables given below.

<i>t</i>	.08	.09	.02	.03	.04	.05	.06	.07	.08	.09
-3.4	.0001	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003
-3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0004
-3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
-3.1	.0010	.0009	.0009	.0008	.0008	.0008	.0008	.0008	.0007	.0007
-3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
-2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0015	.0015
-2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
-2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
-2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
-2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
-2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
-2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
-2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
-2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
-2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
-1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
-1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294

<i>t</i>	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
1.8	.9541	.9549	.9556	.9564	.9571	.9578	.9585	.9593	.9609	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808		
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9845	.9850		

|| testmet.com is sharing your screen.

 Stop sharing Hide



Question 8 Read Later

Select an option

 Oral Response

Consider the following table and answer the questions.

Sheet below shows the data of school boys and girls of an Indian state who reported the number of times they play games, i.e., whether they play every day, never, once a month or once a week.

	Everyday	Never	Once a month	Once a week	Total
Boy	324	154	152	180	650
Girl	216	118	206	168	608

Answer what percentage of girls play at least once a week over the total no. of girls?

 2% 31% 33% 61%

Question 5 Read Later

Select an option

Which of the following lines of code will NOT throw an error for the given two lists?

List1 = [1, 2, 3]

List2 = [4, 5, 6]

 List1 + List2 print(list1 + list2)
print(list1.append(list2)) List1 + List2 print(list1 + list2)
print(list1.append(list2))

Question 5 Revisit Later

Select an option

Which of the following lines of code will NOT throw an error for the given two lists?

```
List1 = [1, 2, 3]
```

```
List2 = [4, 5, 6]
```

 List1 * List2 import numpy as np
np.array(List1)+np.array(List2) List1-List2 import pandas as pd
pd.asarray(L1)+pd.asarray(L2)

Attempted: 1/7

Question 2 Revisit Later

Select an option

 Clear Response

Suppose you have two separate data frames, df1 and df2 as given below. These two dataframes are combined in such a way that the resultant dataframe looks as follows:

	Name	Age	Gender	Blood Group	Height
0	Vikas	24	M	A-	172
1	Rahul	25	M	O+	175
2	Ravina	22	F	AB+	169
3	Viram	31	M	B-	170

These two dataframes are combined in such a way that the resultant dataframe looks as follows:

	Name	Age	Gender	Blood Group	Height
0	Vikas	24	M	A-	172
1	Rahul	25	M	O+	175
2	Ravina	22	F	AB+	169
3	Viram	31	M	B-	170

Which of the following command was used to do this?

`df1.append(df2, axis = 0)``df1.append(df2, axis = 1)``pd.concat([df1, df2], axis = 0)``pd.concat([df1, df2], axis = 1)`

Chat Now

Question 1 Revisit Later**Select an option**

Suppose you have the following two pandas series:

You wish to get the elements in S1 that aren't present in S2.

```
1 S1 = pd.Series([0, 2, 4, 6, 8, 10, 12])  
2 S2 = pd.Series([0, 3, 6, 9, 12, 15])
```

Which of the following command will achieve this?

 S1[S1.isin(S2)] S1[~S1.isin(S2)] S1[S1.isin(~S2)] S2[S1.isin(~S1)]

Attempted: 0/3

Question 1 Revisit Later

Select an option

 Clear Response**What will be the output of the following code?**

```
1 stmt = ['this','is','a','python','test']
2 '-'.join(stmt)
```

this is a python test

this-is-a-python-test

-this is a python test

-thisisapthonstest

Chat Now



Attempted: 0/7

Question 1 Revisit Later**Select an option**

Suppose you have the following two pandas series:

You wish to get the elements in S1 that aren't present in S2.

```
1 S1 = pd.Series([0, 2, 4, 6, 8, 10, 12])  
2 S2 = pd.Series([0, 3, 6, 9, 12, 15])
```

Which of the following command will achieve this?

 S1[S1.isin(S2)] S1[~S1.isin(S2)] S1[S1.isin(~S2)] S2[S1.isin(~S1)]

Keep!!**Question 1****Ansatz****Sigma****Pattern**

Suppose you have the following two partitions:

You want to get the elements of 11 that are present in 12

11 = $\{1, 1, 1, 1, 1, 1, 1, 11\}$

12 = $\{1, 1, 1, 1, 12, 15\}$

What is the following command in R? **ansatz**

ansatz**ansatz****ansatz****ansatz**

Question 2

 Random Selection Clear Response

What will be the output of the following program list?

```
l = [1, 2, 3, 4]
l.append(5)
print(l)
```

 [1, 2, 3, 4] [1, 2, 3, 4, 5, 6, 7, 8, 9] [1, 2, 3, 4, 5, 6, 7, 8]

Done

What will be the output of the following program be?

```
, 2, 3, 4]
cube_func(n):
    n**3
list(map(cube_func, l)))
```

[1, 8, 27, 64]

[1, 2, 3, 4, 1, 2, 3, 4, 1, 2, 3, 4]

[[1, 2, 3, 4], [1, 2, 3, 4], [1, 2, 3,

✓ Error

Attempted 1/1

Question 2

 Revisit Later

Select an option

What will the output of the following program be?

{1, 8, 27, 64}

```
1 l = [1, 2, 3, 4]
2 def cube_func(n):
3     return n**3
4 print(list(map(cube_func, l)))
```

{[1, 2, 4], [2, 3, 4], [2, 16]}

{[1, 2, 4], [1, 2, 4], [1, 2, 4]}

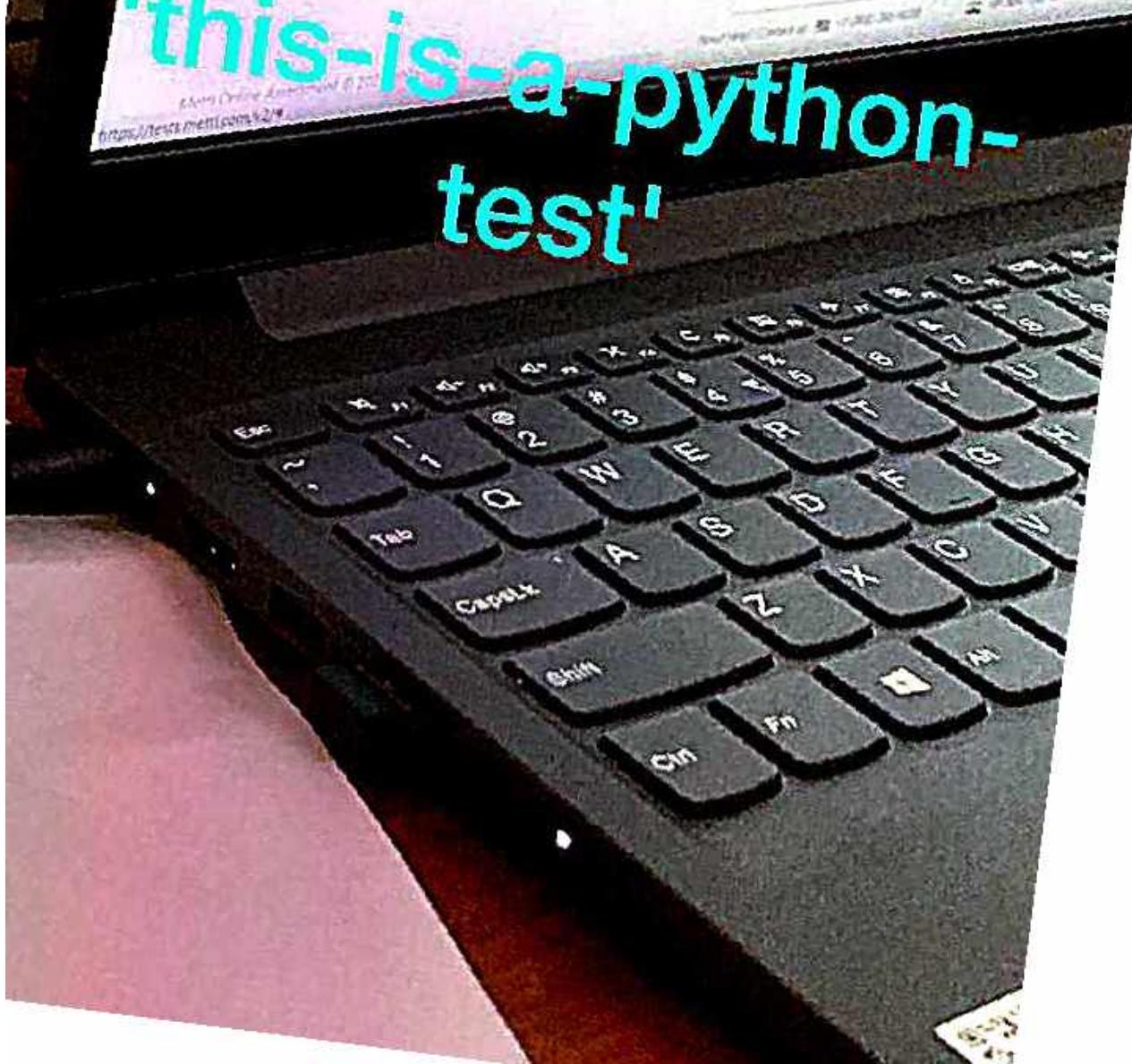
Prev

Question 1

What will be the output of the following code?

```
1 stmt = ['this', 'is', 'a', 'python', 'test']
2 '-' .join(stmt)
```

'this-is-a-python-test'



Question 2

Which of the following code snippets will NOT give the output as [1, 3, 5, 7, 9]?

 Submit Later

Select an option

 Clear Response def printList():
 result = [1, 3, 5, 7, 9] result = [1, 3, 5, 7, 9]
print(result) [1, 3, 5, 7, 9] chars =()
for char in 'Java':
 chars.append(char)
print(chars) None of the above

Re "DS C37 Course 1 Exam - Slot 2" [Inbox](#) 

9:02 AM (9 hours)

Can't read or see images? [View this email in a browser](#)

Dear Candidate ,

You have been invited to take the assessment **DS C37 Course 1 Exam - Slot 2**

The duration of this test is **90 Minutes**.

Please click on the button given below to check your system compatibility before starting the test.

[Check System Compatibility](#)

Please click on the button given below to start the test.

[START TEST](#)

Alternatively, you may copy and paste the below test link in the address bar of a web

Question 1

Q 1 2 3

Answered

Unfinished

5

What will be the output of the following code?

```
2 stat = ['this', 'is', 'a', 'python', 'test']
2 '-'.join(stat)
```

https://www.w3schools.com/python/python_ref_string.asp

Home Office Assessment Platform
https://assessments.mysafespace.com/124



dept_id	last_name	salary	lead_salary
45	Erickson	42000	54000
45	Sutherland	54000	57500
30	Parker	57500	65000
30	Gates	65000	80000
45	Yates	80000	NULL



Rahul Singh 20h

Some people spend their lives trying to create the perfect conditions to live, without really living.

Too many of us believe happiness is a future event. And before we arrive, we need more money first, have a successful career, find a partner, settle down. And only then we will arrive at the destination of happiness. But when we arrive, we will realize happiness isn't there.

Happiness is not found at the finish line. There isn't even a finish line. Life is not a race to be finished; it's a dance to be danced. And only if we allow ourselves to enjoy the dance, can we let happiness in.

One day your life will flash before your eyes and you don't want to see a slide show of all the things that turned out to be irrelevant in your life. Life is happening right now. We've got one shot. Taste the thrill of life. Have the full experience.

The point of living isn't to arrive at the future; it's to arise in the present.

41

@purposologist

@thegoodquote



Send message...





Hi Rohit



nandan@edureka.co ▾

Enter your password

1 Your password was changed 20 hours ago

Show password

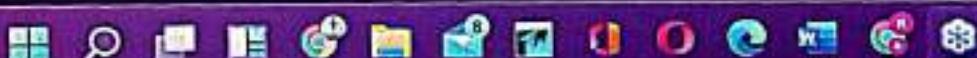
[Forgot password?](#)

[Next](#)



Highlights

- Venturing into Offline Business.
- Catering to Undergrad Student segment primarily.
- Training through Retail Channel Partners & Colleges.
- Blended (Online+Offline) Learning Pedagogy.
- Increasing the employability quotient of colleges students.
- Help colleges improve their campus placements.
- Formal launch in March 2022.



Human Resources - Roadmap

- We are growing and expanding our reach.
- Key focus area → prioritise current and forecast demand to meet business objectives
- Design conducive policies and drive employee friendly work culture
- Employee engagement & recognition

Procure	Develop	Compensate	Integrate
Job Analysis Recruitment Selection Placement Onboarding Transfer Promotion	Performance Appraisal Training Career Planning Development Transition Planning	Evaluation Wages & Salary Bonus & Incentives Payroll	Labor Relations Motivation Grievance Discipline

Enterprise Business – Q4 Initiatives

Key Initiatives:

- Achieved 93L revenue and 1.7 Cr collections in Jan 2022
 - Focus on 3 Cr collections for Q4 and close all overdue payments from clients.
 - Working with BOA and Salesforce to get yearly calendar. First round of meeting is completed. Expecting the update by mid of Feb
 - 7 open positions in Q3/Q4. Delay in hiring and onboarding of resources. Working with HR to expedite hiring
 - New discussions with TCS US for process training requirements –PSP & PMP
 - 2 new client acquisitions in Q4 with an annual business value of 50L each
 - Grow 2 small existing accounts (less than 25L) to grow it to more than 50L+(medium accounts) annually
-

Order Booking	7.2	3.3
Revenue	7.3	2.9
Collection	6.9	2.9

Key Insights:

- Renewed TCS contract for FY23 for a value of 3 Cr
- Closed a new business of 25L from Virtusa
- Won one new account mPokket with a business value of 15L
- Shortlisted for next round of discussion with Soc Gen after submitting RFP. Business Value -25L
- Top 2 contender among 17 vendors for RFP with iQuanti. Business value – 70L annually
- A large potential account Bosch got pushed to FY23. Will re-initiate the discussion in Mar
- No revenue from online leads so far in Q3. 3 potentials in final discussion.



Content

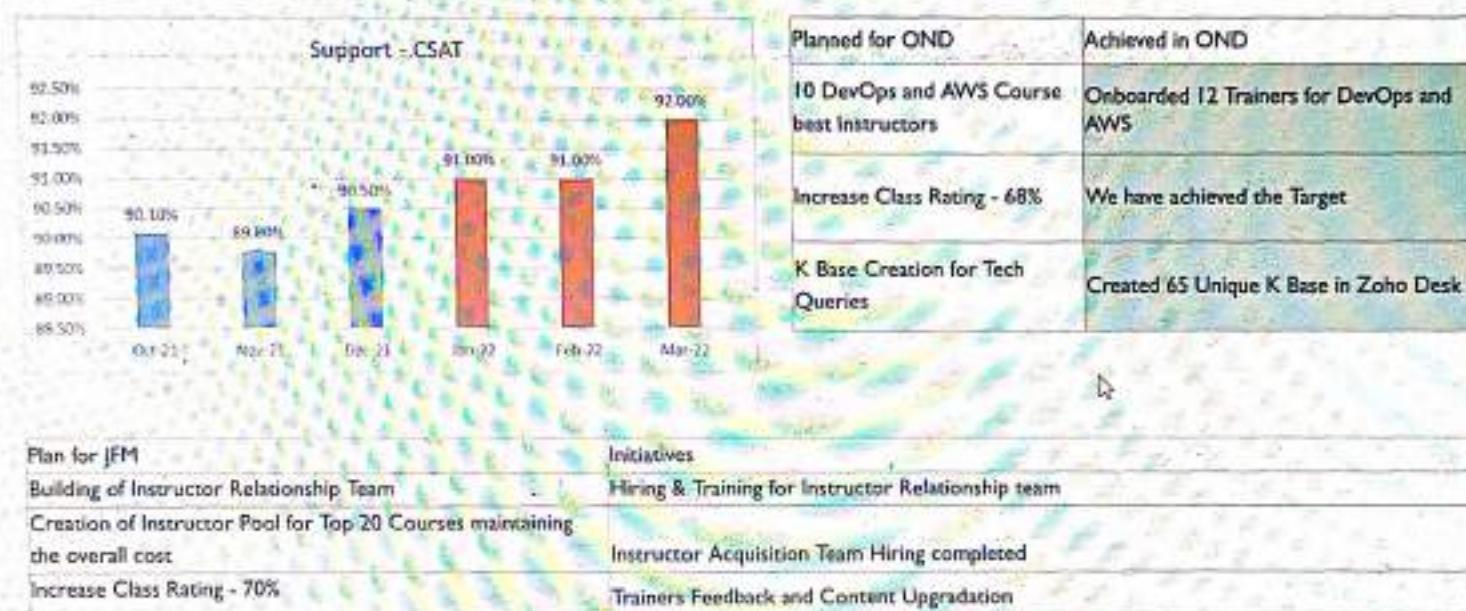


Planned for OND	Achieved in OND
Revamp of old courses	Old courses updated with new and updated Content
Launch of DevOps	New & Updated curriculum for DevOps with Purdue Univ.

Plan for jFM
Improvement of Content Rating
Acacia Content Development
Launch of Univ. associated Program
New Program Launches

Initiatives
Complete OKR based overhaul and org building
Multi-pod Development Methodology
Set up of market research subcommittee
Masters Program on Salesforce | AWS

Operations | Support



Program Management



Planned for Q1D	Achieved in Q1D
Organizational Restructuring	Organization Restructured to Functional Responsibilities
Bring Process Framework into Operations	Created 300+ process documents to streamline work

Plan for JFM
Acacia Program Management
Increase NPS to 45%
Re-build the Edureka Placement Engine
Increase Overall Brand Perception

Initiatives
Hiring & Training for Program Management Team for Acacia
Deep Product intervention into all round Delivery
Placement Assistance & Career Services(PACS) Team creation
Focus on building Customer Success Stories

Sales B2C



Planned for Q1D	Achieved in Q1D
Setup Account management	A team to upgrade existing customer has been setup
Increased Ticket Size	Initiative to sell higher ticket size courses to improve overall avg. ticket has started showing results
Sales Enablement Team	Sales training & refresher team has been setup which will help us to improve conv.
Re-activation Team	This initiative was to re-activate dead leads. We have tried and have paused this initiative for now.

Plan for JFM

Increased revenue from existing customer base

Increase ticket size to 22,200 by Mar'22

Build a scalable team to handle higher lead volumes & improved conversion

Improved People Happiness(Score: 7.74 to 8.13)

Initiatives

- Scale account management team (Hiring, Optimised workflow)
- Set-up referral team
- Focus on increasing mix of higher ticket size courses through Continuous monitoring & training
- Hire 20 more people to maintain optimal lead per agent ratio
- Re-org sales team structure basis country & category
- Agent Readiness (Quicker TAT on Training, Quality Audit)
- Actions based on people survey

Marketing - B2C



Planned for OND	Achieved in OND
Youtube Growth	Reached 3 Million Subscriber Milestone
Blog and Community Growth	Blog remained flat in OND and drop in community traffic
Webinar & Workshop	Doubled the count of webinars from 30 to 60 per month to capture more organic audience

Plan for JFM	Initiatives
Traffic and Lead Growth	<ul style="list-style-type: none"> Bring Top 20 courses in Rank 1-5 Create Freemium funnel to grow organic leads and bring down overall CPL Optimize Paid Campaigns to reduce CPL to 1470 (Overall CPL from 650 to 600) Establish new affiliate partners/channels to grow affiliate leads and revenue
Improve Brand Traffic	<ul style="list-style-type: none"> Increase followers on different social media platforms Bring rating to 4.5 and above across all review platforms
Hiring	<ul style="list-style-type: none"> Fill the hiring requirements of content marketing, SEO and Paid marketing teams

OverAll Business (B2C | PGP)

OND - 98% Achieved

Jan - 101% Achieved

Planned for OND	Achieved in OND
Industry Tie-Ups	Microsoft tie up is done. IBM is still pending.
Launch New Programs	Cyber Security Masters, DevOps PGP Program with Purdue Univ were launched
Define Team wise OKR	Breaking the business level objectives into team level objectives & initiatives

Plan for JFM	Initiatives
Grow Non-Paid Funnel	Increase revenue from non-paid sources like SEO, Freemium Funnel, Referral & Repeat
Increase Revenue per lead (Conv. & Ticket Size)	Focus on adding industry tie-ups, train & re-train sales agents, sell higher ticket size programs
Launch New Programs	Launch CEH, Salesforce masters, AWS masters and an Advanced Certification Program
Initiate Branding Activities	Start sending Customer Kit & Instructor Kit & Focus on Brand perception

90 min
left

1. Python

Which of the following is the correct output of this program?

```
num1 = [4, 5, 6]
num2 = [5, 6, 7]
result = map(lambda n1, n2: n1+n2, num1, num2)
print(list(result))
```

Answer Options

Select any one option

[9, 11, 13]

[3, 2, 1]

[4, 5, 6]

Invalid syntax

1/34

>>

Answered



Submit

89 min
left

3. Python

If you have a list L = [3, 1, 8, 9, 7], what will the list look like when you perform L.insert(2, 4)?

Answer Options

Select any one option

[3, 1, 8, 9, 2, 7]

2/34 >> [3, 1, 4, 8, 9, 7]
swered

[3, 1, 8, 9, 7, 2, 4]

[2, 4, 3, 1, 8, 9, 7]

4. Python for DS

◀ Previous

Consider a data frame df that consists of the column 'Name'. Assume that all the names with the last two letters 'ic' are from Croatia, and otherwise the names are from another country. Choose the code to create a new column called 'Nationality'. This column must contain 'Croatian' when the name is from Croatia and 'Others' when the name is from other countries.

Sample Input:

	Name
0	Modric
1	Rakitic
2	Ronaldo

Sample output:

	Name	Nationality
0	Modric	Croatian
1	Rakitic	Croatian
2	Ronaldo	Others



Answer Options

Select any one option

- df['Nationality'] = df['Name'].apply(lambda x: 'Croatian' if x[-2:]=='ic' else 'Others')
- df['Name'].apply(lambda x: 'Croatian' if x[-1:] else 'Others')
- df['Nationality'] = df['Name'].apply(lambda x: 'Croatian' if x[-2:]=='ic' else 'Others')
- df['Name'].apply(lambda x: 'Croatian' if x[-2:-1]=='ic' else 'Others')



89 min
left

2. Python

What is the output of the following program?

```
print((2, 5) + (5, 2))
```

Answer Options

Select any one option

(2, 5), (5, 2)

>>

(7, 7)

Answered

 (2, 5, 5, 2)

Invalid syntax



:



Submit

85 min
left

6. Python for DS

Which of the following code snippets is used to find the sum of the elements of a NumPy array 'arr'?

Answer Options

Select any one option

arr.sum()



5/34
Answered

arr.add()

add(arr)

arr.sum



⋮



Submit

84 min left

7. Python for DS

MCQs

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8
- 9

- 10
- 11
- 12
- 13
- 14
- 15
- 16
- 17
- 18
- 19
- 20
- 21
- 22
- 23
- 24
- 25
- 26
- 27

Suppose you have the following two dataframes df1 & df2:

	X1	X2		X1	X3
0	A	1		0	A I
1	B	2		1	B II
2	C	3		2	D IV

You combined these dataframes such that the resultant dataframe looks like the following:

	X1	X2	X3
0	A	1.0	I
1	B	2.0	II
2	C	3.0	NaN
3	D	NaN	IV

Which command was used to achieve this?

 Answer Options

Select any one option

pd.merge(df1, df2, how='left', on='X1')

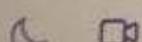
2 3

pd.merge(df1, df2, how='right', on='X1')



pd.merge(df1, df2, how='inner', on='X1')

pd.merge(df1, df2, how='outer', on='X1')



 Submit Test

82 min left

10. Data Visualisation in Python

Which of the following data summary is not included in box plot?

- 2
- 3
- 5
- 6
- 8
- 9
- 11
- 12
- 14
- 15
- 17
- 18

Answer Options

Select any one option

Mean

Quartile

Count

Maximum

82 min left

9. Python for DS

MCQs

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8
- 9
- 10
- 11
- 12
- 13
- 14
- 15
- 16
- 17
- 18
- 19
- 20
- 21
- 22
- 23
- 24
- 25
- 26
- 27
- 28
- 29
- 30

Fill in the blank in the following sentence

Consider a data frame of containing a column with the name 'Profit'. The output of df['Profit'] is a

Answer Options

Select any one option

Pandas Series

Python list

NumPy array

Python dictionary

Coding

- 1
- 2
- 3

4



Submit Test



Answer Options

Select any one option

- df.distplot('Cart Values', bins = 20, kde = False)
- df.sns.distplot('Cart Values', bins = 20, kde = False)

sns.distplot('Cart Values', bins = 20, kde = False)

- df.distplot('Cart Values', bins = 20, kde = True)

11. Data Visualisation in Python

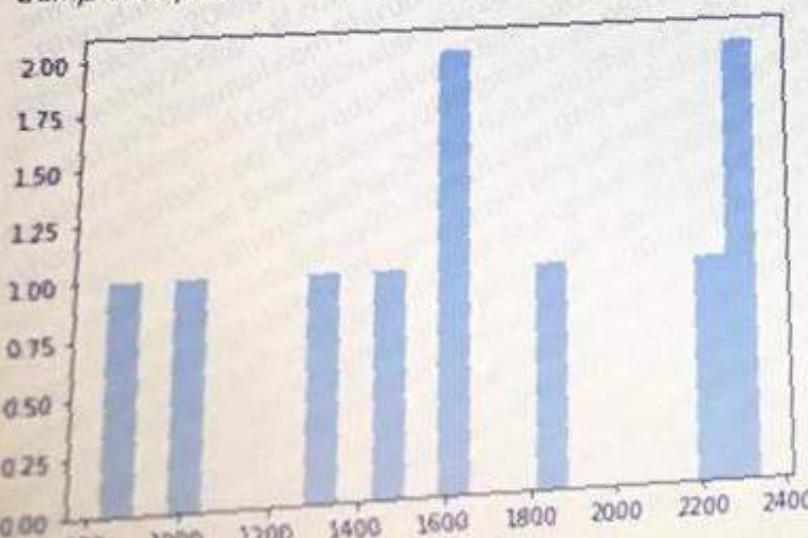
[Previous](#)[Next >](#)

An e-commerce website wants to reduce the churn rate by offering discounts on specific cart values. For this purpose, you want to divide cart values into 20 ranges and display the frequency of each range. Which of the following code snippets will do this task?

Sample input:

Cart Values
0 1589
1 827
2 1644
3 1480
4 2309
5 1033
6 1820
7 2249
8 2342
9 1332

Sample output:



12. Exploratory Data Analysis

Suppose you are given data in which date is a column with the format dd-mm-yyyy. You have observed that both month and year are the same across all the rows. Which of the below operations help in the day wise analysis of the data?

Answer Options

Select any one option

Clear Answer

Skip dd and make two columns for mm and yyyy.

Skip dd and create a column with mm-yyyy.

Skip both mm and yyyy and create a column with dd.

All of the above



13. Exploratory Data Analysis

X. Previous

Imagine you have two numerical variables, X and Y. You want to predict the value of Y based on the value of X. You build a scatterplot to understand the relationship between the two variables. What information can you gather from the given graph?

Predicting Y



Answer Options

3

Select any one option

- Variables have a strong positive correlation
- Variables have a strong negative correlation
- Variables have a weak positive correlation
- Variables have no correlation



15. Inferential Statistics

< Previous Next >

Suppose X is a normally distributed variable with a mean (μ) of 43 and a standard deviation (σ) of 6.4. Determine the probability of $X < 32$, given that $P(z < -1.71) = 0.341$.

Answer Options

Select any one option

Clear Answer

0.341

0.962

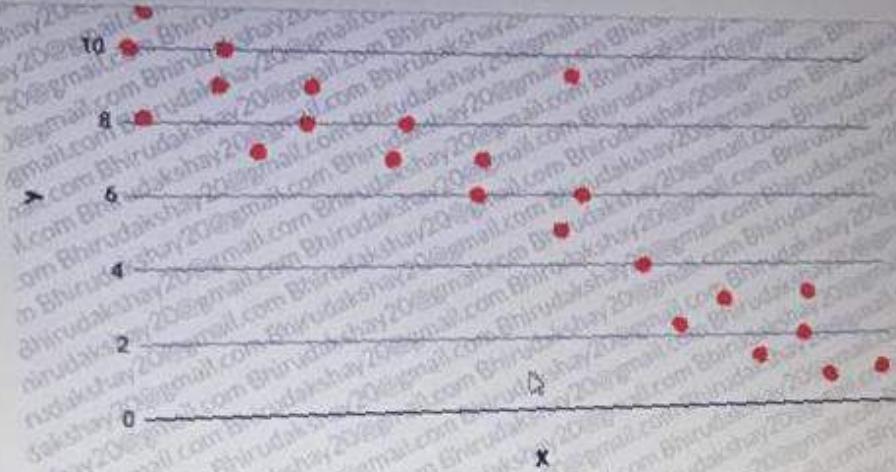
6.231

0.44



13. Exploratory Data Analysis

< Prev



Answer Options

Select any one option

Variables have a strong positive correlation

Variables have a strong negative correlation

Variables have a weak positive correlation

Variables have no correlation

3

nit Test



16. Inferential Statistics

[◀ Previous](#)[Next ▶](#)

The duration between the charges of the battery of a personal computer is normally distributed with a mean of 66 hours and a standard deviation of 20 hours. What is the probability that the duration will be between 58 and 75 hours, given that $P(z < 0.45) - P(z < -0.4) = 0.0443$?

Answer Options

Select any one option

[Clear Answer](#) 0.595 3.44 0.0443 1.98

The probability of a person having a mosquito bite is $P(B) = 0.8$, and the probability of a person having disease is $P(A) = 0.4$. If the probability of a person getting a mosquito bite, given that the person has the disease is $P(B|A) = 0.6$, find the probability of a person having the disease after they get a mosquito bite.

Answer Options

Select any one option

Clear Answer

 0.2 0.3 0.7 0.9

17. Inferential Statistics

Consider the following discrete probability distribution.

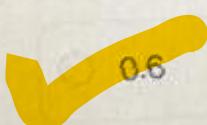
x	p(x)
2	0.1
4	?
6	0.2
8	0.4

Find the probability that x is greater than 4.

Answer Options

Select any one option

0.2

 0.6

0.8

0.1



19. Inferential Statistics

For a normal distribution, which of the following does **not** coincide with the mean?

3

6

9

12

15

18

21

24

27



Mode

30



Standard deviation

Both a and b

3

21. Hypothesis Testing

◀ Previous Next ▶

The amount of a certain trace element in blood is known to vary with a standard deviation of 14.1 ppm (parts per million) for male blood donors and 9.5 ppm for female donors. Random samples of 75 male and 50 female donors yield concentration means of 28 and 33 ppm, respectively. Find the Z-score?

Answer Options

Select any one option

Clear Answer

-2.34



-2.73

 -2.37

+2.37



23. Hypothesis testing

< Previous Next >

A chewing gum company claims that each chewing gum unit contains more than 2% Glycerine by mass on average. Assuming that you want to do a hypothesis test to test this claim, which of the following is true?

Answer Options

Select any one option

The null hypothesis is that the average Glycerine content is not equal to 2%.

The alternate hypothesis is that the average Glycerine content is less than 2%.

The null hypothesis is that the average Glycerine content is less than or equal to 2%.

The alternate hypothesis is that the Glycerine content is equal to 2%.



22. Hypothesis testing

As the p-value increases, the chance of null hypothesis getting rejected:

Answer Options

Select any one option

Increases

 Decreases

Remains the same

Cannot be determined

24. SQL

From the options below, select the correct statement about the **DROP** clause or the **TRUNCATE** clause.

Answer Options

Select any one option

- The **DROP** command is used to delete all the rows from a table
- TRUNCATE removes the schema of a table.
- The **DROP** command removes the schema of a table as well as the table's contents
- The **TRUNCATE** command removes selected rows from a table

26. SQL

What does the following code snippet do?

```
ALTER TABLE STUDENT ADD(MARKS VARCHAR(20));
```

Answer Options

Select any one option

Adds a column called MARKS in the table STUDENT

Checks if a column called MARKS is present in the table STUDENT

Invalid Syntax

None of the above

25. SQL

Select the correct statement from the options given below:

3

6

9

12

15

18

21

24

27

30

3

Answer Options

Select any one option

CREATE and ALTER are DML statements.

ALTER and DROP are DDL statements.

UPDATE and DELETE are DDL statements.

INSERT and ALTER are DML statements.

28. Advanced SQL

[Previous](#) [Next](#)

Akshay is on the look for a new mobile phone and has scrapped the mobile phone data from the Flipkart website. The scraped table 'data' has the following columns.

serial_number	manufacturer	model	price	number_of_ratings	average_rating	median_rating
---------------	--------------	-------	-------	-------------------	----------------	---------------

Akshay wants a phone for an amount less than 30000 which has at least 5000 ratings with an average rating of at least 4.2.

Select the correct query to print the top 3 phones fulfilling the conditions set up by Akshay. The top three phones have to be selected based on the median_rating.

Answer Options

Select any one option

[Clear Answer](#)

select *,
rank() over (order by median_rating desc) as 'rank'
from data
where price < 30000 and number_of_ratings > 4999 and average_rating > 4.2
limit 3;

select *,
rank() over (order by median_rating) as 'rank'
from data
where price < 30000 and number_of_ratings > 4999 and average_rating > 4.2
limit 3;

select *,
rank() over (order by median_rating desc) as 'rank'
from data
where price < 30000 and number_of_ratings > 4999 and average_rating > 4.2 ;

 select *



27. Advanced SQL

[« Previous](#)[Next »](#)

A major supplier of solar panels sells their panels to numerous stores. In the table "Sales_data", data about each sale is noted. In this table, the "Store_id" refers to unique ids given to each store (which are customers of the supplier). Another table called "Customer" holds data for all customer stores. So, "Store_id" and "Customer_id" are the same.

```
SELECT [Location]
FROM Sales_data
WHERE Store_id NOT IN
    (SELECT Customer_ID
     FROM Customer
     WHERE area = 213);
```

Which of the following statements summarizes the output of the query?

« Answer Options

Select any one option

[Clear Answer](#)

- The result table will have the locations of all stores in the 213-area.
- The result table will have "Customer_id" for all customers in the 213 area.
- The result table will have locations of all stores not in the 213 area.

The result table will have "Store_id" of all stores not in the 213 area.



29. Advanced SQL

[◀ Previous](#) [Next ▶](#)

Consider the table 'score' containing batting data in ODIs of the Indian Cricket Team. The table structure is as below

player_id	game_date	score	opposition
-----------	-----------	-------	------------

Please note that all scores are considered as if the player got out. Not outs are not to be considered for the purpose of calculation. You are interested to determine the player who is the best at his peak. For the purpose, you have decided to calculate 5 match moving average and select the player who has the maximum. Which of the following pieces of the code shall you use to achieve the same?

Answer Options

Select any one option

[Clear Answer](#)

- select player_id, game_date,
avg(score) over (partition by player_id order by game_date rows between 4 preceding and 0 following) as moving_average
from score
order by moving_average desc
limit 1;
- select player_id, game_date,
avg(score) over (partition by player_id order by game_date desc rows between 4 preceding and 0 following) as moving_average
from score
order by moving_average desc
limit 1;
- select player_id, game_date,
avg(score) over (partition by player_id order by game_date rows between 4 preceding and 0 following) as moving_average
from score
order by moving_average desc ;
- select player_id, game_date,
avg(score) over (partition by player_id order by game_date desc rows between 4 preceding and 0 following) as moving_average
from score



4 5 6 7 8 9)

30. Advanced SQL

< Previous Next >

Which of the following SQL queries will result in an error message being displayed? Assume that all column names and table names are present in the database.

Answer Options

Select any one option.

Clear Answer

select fname,hno,
rank(order by dno) over (partition by dno order by hno) as 'rank'
from employee;

select fname,hno,
rank() over (partition by dno order by hno) as 'rank'
from employee;

select fname,hno,
rank() over (order by hno) as 'rank'
from employee;

select fname,hno,
rank() over (partition by dno order by hno) as 'rank'
from employee;

Test



A portion of a computer keyboard is visible at the bottom of the screen, showing keys such as #, \$, %, ^, &, *, (,), 0, and various function keys like F1 through F12.

and select the player who has the maximum. Which of the following pieces of the code shall you use to achieve the same?

Answer Options

Select any one option

Clear Answer

select player_id, game_date,
avg(score) over (partition by player_id order by game_date rows between 4 preceding and 0 following) as moving_average
from score
order by moving_average desc
limit 1;

select player_id, game_date,
avg(score) over (partition by player_id order by game_date desc rows between 4 preceding and 0 following) as moving_average
from score
order by moving_average desc
limit 1;

select player_id, game_date,
avg(score) over (partition by player_id order by game_date rows between 4 preceding and 0 following) as moving_average
from score
order by moving_average desc ;

select player_id, game_date,
avg(score) over (partition by player_id order by game_date desc rows between 4 preceding and 0 following) as moving_average
from score
order by moving_average desc ;



1 2 3 4 5 6 7 8 9 0 . , - + =

\$ % ^ & * ') - + =

R T Y U I O P { }

1. String Reverser

[Previous](#)[Next >](#)

Problem Statement

Given a string *s*. Write a program to reverse the order of the words present in it.

Function description

Complete the ***strReverse*** function in the editor below. It has the following parameter(s):

Name	Type	Description
<i>s</i>	STRING	The input string to be reversed

Return The function must return a *STRING* denoting the reverse of the original string provided

Constraints

Input format for debugging

Sample Testcases

3. Salary Analysis

[Previous](#) [Next](#)

You are given a table **EMPLOYEE**. You want to do a comparative analysis of employee salaries using the data in this table.

Write a MySQL query to display the **EMP_NAME**, **EMP_SALARY** and **DEPT_ID** of employees whose **EMP_SALARY** is greater than the average **EMP_SALARY** and they have a **EMP_NO** greater than 103.

Notes:

- It is given that since the schema is defined using a temporary table you are **not allowed** to use queries that try to access the same table **more than once** in a **single query** to compute the final output.

Schema

Table structure

EMPLOYEE

Name	Type	Description
EMP_NO	int	Column denoting EMP_NO representing employee number
EMP_NAME	varchar(50)	Column denoting EMP_NAME representing employee name
HIRE_DATE	date	Column denoting HIRE_DATE representing date on which employee is hired
EMP_SALARY	int	Column denoting EMP_SALARY representing salary of the employee
DEPT_ID	int	Column denoting DEPT_ID representing id of the department where the employee works

Sample testcase 1



2. Unique Students

[Previous](#)[Next](#)

Problem Statement

Given two lists of students who are enrolled in Data Science and Machine Learning course respectively, find the list of students who are enrolled in exactly one program.
Make sure you sort the final list alphabetically.

Function description

Complete the **unique** function in the editor below. It has the following parameter(s):

Name	Type	Description
n	INTEGER	Number of students in Data Science Program
DS	STRING ARRAY	Names of students in Data Science Program
m	INTEGER	Number of students in Machine Learning Program
ML	STRING ARRAY	Names of students in Machine Learning Program
Return	The function must return a STRING ARRAY denoting the Names of student only enrolled in single program and are alphabetically arranged	

Constraints

- $1 \leq n \leq 10^5$
- $1 < \text{len}(DS[i]) \leq 10^5$

3. Salary Analysis

[Previous](#)[Next](#)

Name	Type	Description
EMP_NO	int	Column denoting EMP_NO representing employee number
EMP_NAME	varchar(50)	Column denoting EMP_NAME representing employee name
HIRE_DATE	date	Column denoting HIRE_DATE representing date on which employee is hired
EMP_SALARY	int	Column denoting EMP_SALARY representing salary of the employee
DEPT_ID	int	Column denoting DEPT_ID representing id of the department where the employee works

Sample testcase 1

Input

EMPLOYEE

EMP_NO	EMP_NAME	HIRE_DATE	EMP_SALARY	DEPT_ID
103	Vipul	1990-10-11	5000	34
104	John	2020-11-11	3000	15
105	Ram	2020-10-11	10000	34

Output

Ram

10000

34

4. Latest Hire

[Previous](#)[Next](#)

Schema

```

EMPLOYEE,
CREATE TEMPORARY TABLE `EMPLOYEE` (
    `EMP_NO` int NOT NULL,
    `EMP_NAME` varchar(50) DEFAULT NULL,
    `HIRE_DATE` date DEFAULT NULL,
    `EMP_SALARY` int DEFAULT NULL,
    `DEPT_ID` int DEFAULT NULL,
    PRIMARY KEY (`EMP_NO`)
) ENGINE = InnoDB DEFAULT CHARSET = latin1

```

Table structure

EMPLOYEE

Name	Type	Description
EMP_NO	int	Column denoting EMP_NO representing employee number
EMP_NAME	varchar(50)	Column denoting EMP_NAME representing employee name
HIRE_DATE	date	Column denoting HIRE_DATE representing date on which employee is hired
EMP_SALARY	int	Column denoting EMP_SALARY representing salary of the employee
DEPT_ID	int	Column denoting DEPT_ID representing id of the department where the employee works

Sample testcase 1

You want to find recently hired employee based on **DEPT_ID**.

Write a MySQL query to display all the details of the employees of a department with **DEPT_ID 123** who has been hired on the latest date.

Note:

- It is given that since the schema is defined using a temporary table you are **not allowed** to use queries that try to access the same table **more than once** in a **single query** to compute the final output.

Schema

```
EMPLOYEE,
CREATE TEMPORARY TABLE `EMPLOYEE` (
    `EMP_NO` int NOT NULL,
    `EMP_NAME` varchar(50) DEFAULT NULL,
    `HIRE_DATE` date DEFAULT NULL,
    `EMP_SALARY` int DEFAULT NULL,
    `DEPT_ID` int DEFAULT NULL,
    PRIMARY KEY (`EMP_NO`)
) ENGINE = InnoDB DEFAULT CHARSET = latin1
```

Table structure

EMPLOYEE

Name	Type	Description
EMP_NO	int	Column denoting EMP_NO representing employee number
EMP_NAME	varchar(50)	Column denoting EMP_NAME representing employee name
HIRE_DATE	date	Column denoting HIRE_DATE representing date on which employee is hired

4. Latest Hire

[Previous](#)[Next](#)

Name	Type	Description
EMP_NO	int	Column denoting EMP_NO representing employee number
EMP_NAME	varchar(50)	Column denoting EMP_NAME representing employee name
HIRE_DATE	date	Column denoting HIRE_DATE representing date on which employee is hired
EMP_SALARY	int	Column denoting EMP_SALARY representing salary of the employee
DEPT_ID	int	Column denoting DEPT_ID representing id of the department where the employee works

Sample testcase 1

Input

EMPLOYEE

EMP_NO	EMP_NAME	HIRE_DATE	EMP_SALARY	DEPT_ID
101	Ram	2020-10-11	1000	34
102	John	2020-11-11	4500	123
103	Vipul	1990-10-11	90	34

Output

102	John	2020-11-11	4500	123
-----	------	------------	------	-----



```
7]: def unique(n,DS,m,ML):
    DS_set = set(DS)
    ML_set = set(ML)

    unique = sorted(DS_set.symmetric_difference(ML_set))

    string = "\n".join(unique)

    return string
```

```
def strReverse(s):  
    words = s.split(' ')  
    string = []  
    for word in words:  
        string.insert(0, word)  
    result = " ".join(string)  
    return result
```

```
~~~~~  
  
def main():  
    s = sys.stdin.readline().str  
  
    result = strReverse(s)  
  
    print(result)  
  
if __name__ == "__main__":  
    main()
```

```
3 Enter your query here
2
3 - Notes: MySQL queries are case-sensitive.
4
5 SELECT * FROM EMPLOYEE
   WHERE DEPT_ID = 123
   ORDER BY HIRE_DATE DESC;
```

thon 3

```
1 def unique(n,DS,m,ML):
2     DS_set = set(DS)
3     ML_set = set(ML)
4
5     unique = sorted(DS_set.symmetric_difference(ML_set))
6
7     string = "\n".join(unique)
8     return string
9
10
11 def main():
12
13     n = int(input())
14     DS=[None]*n
15     for j in range(n):
16         DS[j] = input()
17
18
19     m = int(input())
20     ML=[None]*m
21     for j in range(m):
22         ML[j] = input()
23
24     result = unique(n,DS,m,ML);
25     print(result)
26 if __name__ == "__main__":
27     main()
```

Console Custom Test Case



Run code

All Test cases passed

✓ Test Case #1	Accepted	memory 9992kb
✓ Test Case #2	Accepted	memory 9996kb
✓ Test Case #3	Accepted	memory 9992kb

14. Exploratory Data Analysis

Which of the following actions is **not** associated with data cleaning?

Answer Options

Select any one option

- Dropping unnecessary header rows from a table
- Checking for irregularities in data types in a column
- Removing unnecessary rows and columns
- Creating a pair plot to compare two variables