26. C2 Business Problem Solving

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 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

- 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

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

- 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

Submit Test

27. C2 linear Regression

< Previous Next >

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.

Answer Options

Select any one option

Clear Answer

- 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

Logistic Regression

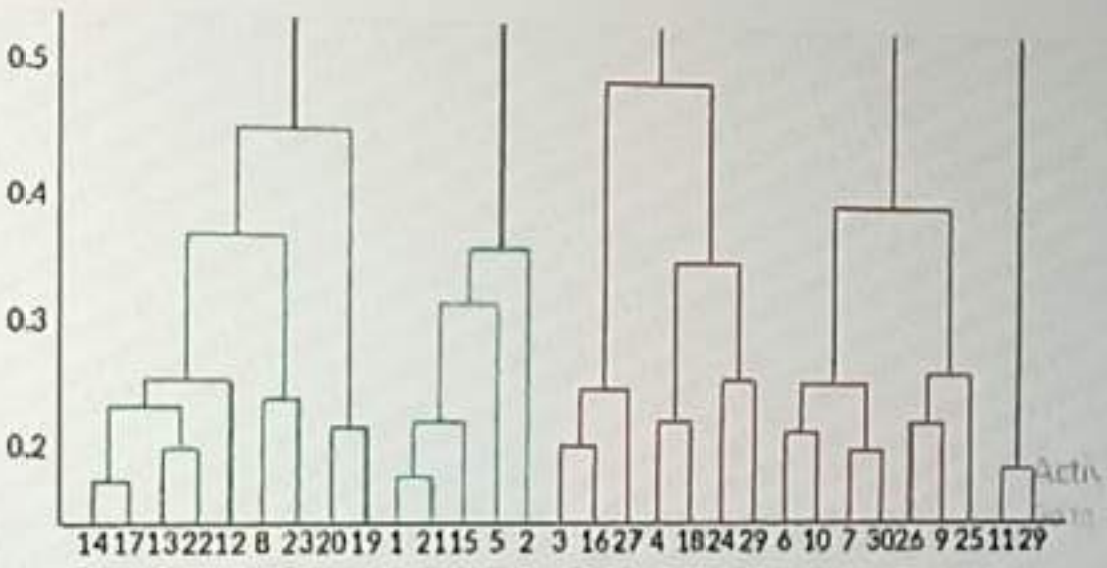
< Prev

e built a Logistic Regression model that is trying to predict whether a loan is approved or not based on a person's F
model parameters: Intercept (B0) = -9.346 and coefficient of FICO score = 0.0146. Given these parameters, can you
ty of a loan getting approved for someone with a FICO score of 655?

ptions

ne option

28. C2 Clustering



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 interpretation is not possible for K-Means clustering

97 min left

29. 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
28 29 30
31 32 33
34 35 36
37

Course 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 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

Clear Answer

Select any one option

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

124 min left

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

18. C2 Logistic Regression

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

Answer Options

Select any one option

- 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.

min left

30. C2 linear Regression

‹ Previous Next ›

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 Answer

A, B, C and D

A, C, and D

A, B and C

B, C and D

2 3
5 6
8 9
11 12
14 15
17 18
20 21
23 24
26 27
29 30
32 33
35 36

Submit Test

min left

31. C2 Multiple options correct

Previous Next

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 Answer

- [x] Silhouette Score
- [x] Elbow Curve
- [ ] Hopkins Statistic
- [ ] Dendogram

2 3
5 6
8 9
11 12
14 15
17 18
20 21
23 24
26 27
29 30
32 33
35 36
2 3

Submit Test

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.)

swer Options

ct one or more options

Clear Answe

The area under the ROC curve can be more than 1.

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

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

uments passed in the above function are actual values of the target variable and the respective predicte

23. C2 Multiple correct answer

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 Answer

- [x]
```
from sklearn.linear_model import LogisticRegression
lr = LogisticRegression()
lr.fit(X_train, y_train)
```

- [x]
```
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()
```

How is regression different from classification?

Answer Options

Select any one option

- 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

Select one or more options

Clear

- [x] The learning rate of curve C is highest among all curves
- [ ] The learning rate for curve B is lower than A
- [x] 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.

29. C2 Clustering

Previous

Silhouette metric for any ith 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

- b(i) is the average distance from the nearest neighbour cluster(Separation).
- a(i) is the average distance from own cluster(Cohesion).
- a(i) is the average distance from own cluster(Cohesion).
- Silhouette Metric ranges from 0 to +1

Previous

Which of the following statements is NOT true?

Answer Options

Select any one option

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

35. C2 linear Regression

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 regress model)

Answer Options

Select any one option

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)

36. C2-Basics of NLP and Text Mining

2
2 3
5 6
8 9
11 12
14 15
17 18
20 21
23 24
26 27
29 30
32 33
35 36
3
2 3

Which of the following strings will match with the regular expression "01*0$?
1. 0
2. 00
3. 01111111110

Answer Options

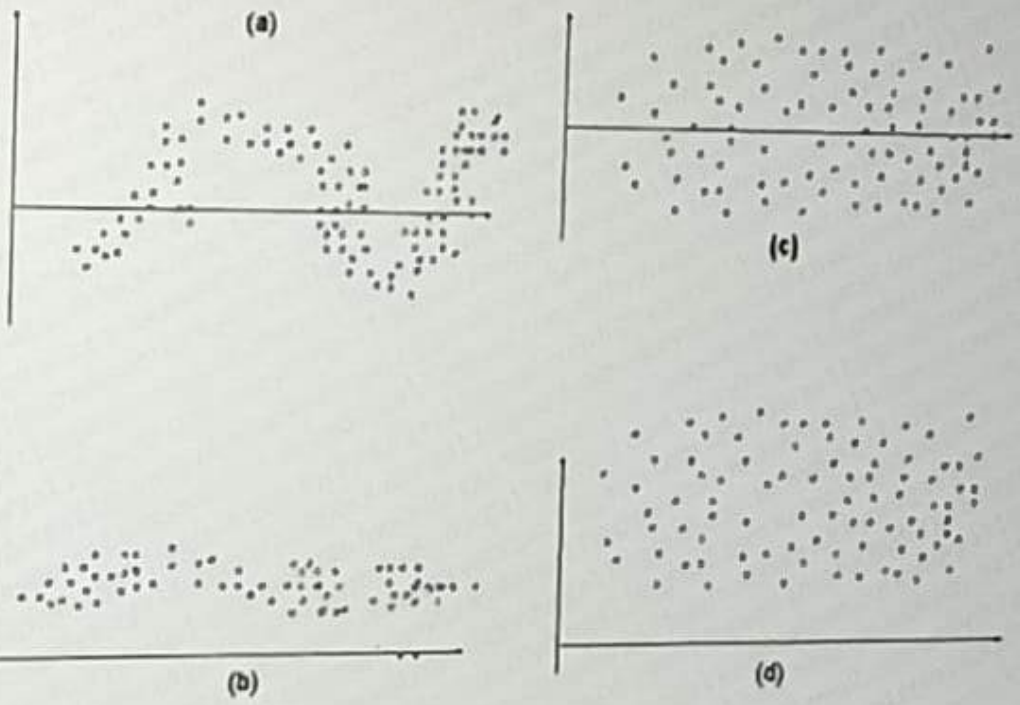
Select any one option

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

left

37. C2 linear Regression

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

A

B

C

D

Test

Answer Options

Select any one option

- CREATE TABLE MyGuests (id INT(6) UNSIGNED AUTO_INCREMENT PRIMARY KEY,firstname VARCHAR(30)
- INSERT INTO MyGuests (firstname, lastname, email) VALUES ('John', 'Doe', 'john@example.c
- UPDATE MyGuests SET lastname='Doe' WHERE id=2
- DELETE FROM MyGuests WHERE id=3

3. C3 Intro to Cloud and AWS

Suppose you have been using services of a cloud service provider for a few years, and now you want to move your current cloud infrastructure from the present cloud service provider to another. Which of the following characteristics of cloud allows you to do so efficiently and cost-effectively?

Answer Options

Select any one option

- Multi-tenancy
- On-Demand Self-Service
- Infrastructure as Code(IaC)
- Rapid Elasticity

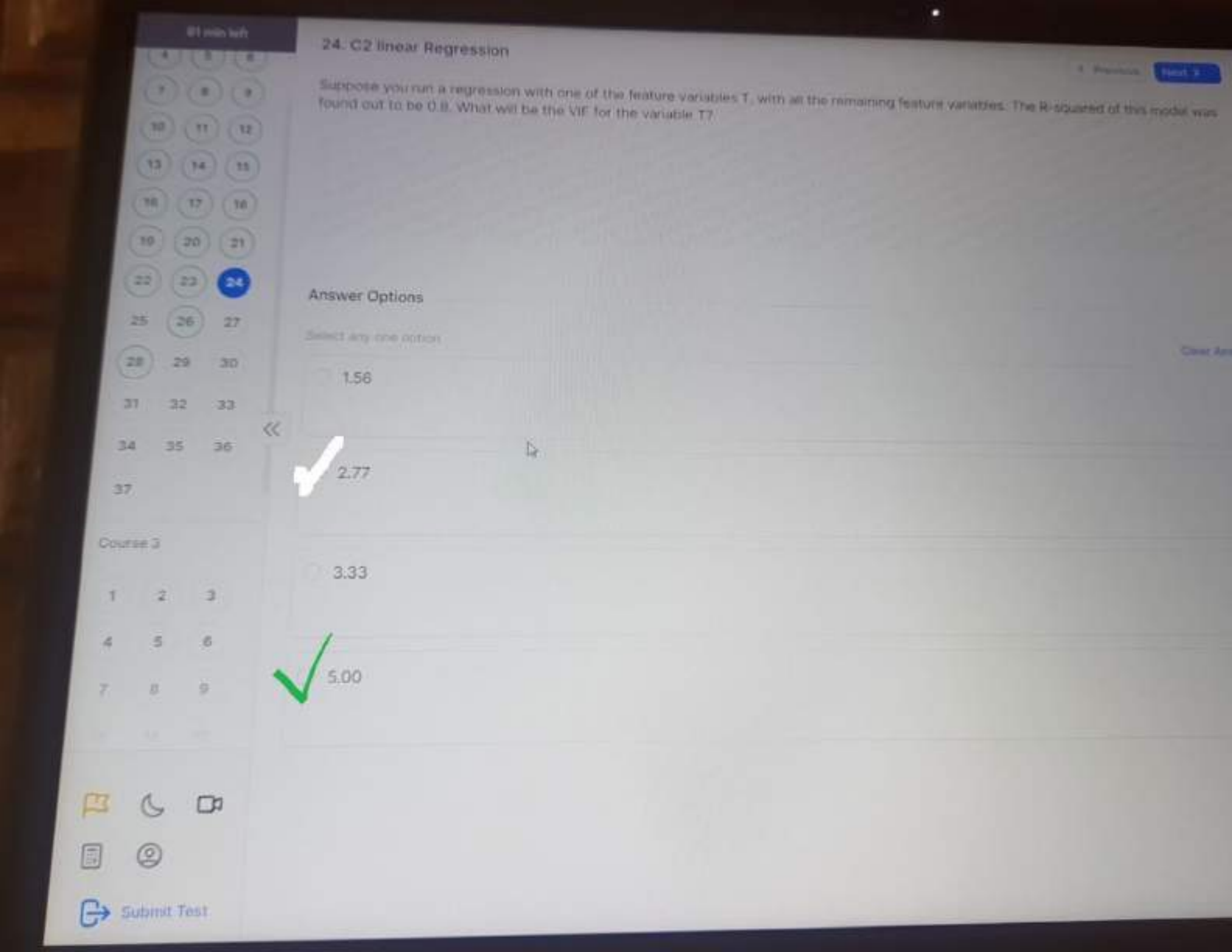
Silhouette metric for any ith point is given by: S(i)=(b(i)-a(i))/max(b(i),a(i))
Which of the following is not true about Silhouette metric?

Answer Options

Select any one option

- 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 cluste
- ✓ Silhouette Metric ranges from 0 to +1

Course 3


24. C2 linear Regression
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
1.56
2.77
3.33
5.00
Submit Test

66/66

Answered

Answer Options

Select any one option

- 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
- F1-score is always the best metric for evaluating a logistic re

Submit

66/66

Answered

Answer Options

Select any one option

- 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
- F1-score is always the best metric for evaluating a logistic re

Submit

Answer Options

Select any one option

It helps in finding the different predictive patterns for the different segments

WoE helps in treating missing values for both continuous and categorical

WoE values should follow an increasing or decreasing trend across bi

All of the above

Submit

Answer Options

Select any one option

It helps in finding the different predictive patterns for the different segments

WoE helps in treating missing values for both continuous and categorical

WoE values should follow an increasing or decreasing trend across bi

All of the above

Submit

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

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

ratio of positive to negative class is 1:10000. You run
find out that the model has a high value of precision and a low value of recall. Which of

Answer Options

Select any one option

66/66

Answered

- (●) The class is handled well by the data
- ( ) The model is not able to detect the class, but when it does it is highly trustat
- ( ) The model is able to detect the class but it includes data points from the ot
- ( ) The class is handled poorly by the data

Submit

PaaS model?

Answer Options

Select any one option

- IaaS model offers reduced maintenance from the user end.
- IaaS model usually provides more flexibility in selecting the unde
- IaaS model removes the complexity of setting up, configuring, operating systems.
- All of the above

mit

application that must be run on four EC2 instances. Out of the four EC2 i
execute mission-critical software and need to be run all the time. The third EC2 instar
gets loaded only when the customer accesses it from time to time; customer uptime
Finally, the last EC2 instance runs a background job that collates the logs from time to
Which would be a cost-effective combination of instances for this purpose?

Answer Options

Select any one option

- Three on-demand instances and one spot instance
- One reserved instance with a partial upfront payment and three spot instances
- Two reserved instances with an upfront payment and one on-demand instance
- Four on-demand instances

Which strategy would you use here to store the DataFrame in memory

Answer Options

Select any one option

- Add checkpoints to store DataFrame in HDFS
- Cache the DataFrame that has been used multiple times
- Create temp tables of DataFrame
- Merge all the data frames and combine all the queries in

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

```
raw_roc(actual, probs):
tpr,thresholds=metrics.roc_curve(actual,probs,drop_intermediate=False)
_score = metrics.roc_auc_score(actual, probs)
rn None
```

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

r Options

ne or more options

Clear Answer

e area under the ROC curve can be more than 1.

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

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

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

11. C2 linear Regression

Previous | Next

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 Answer

- 1.56
- 2.77
- 3.33
- 5.0

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

Answer Options

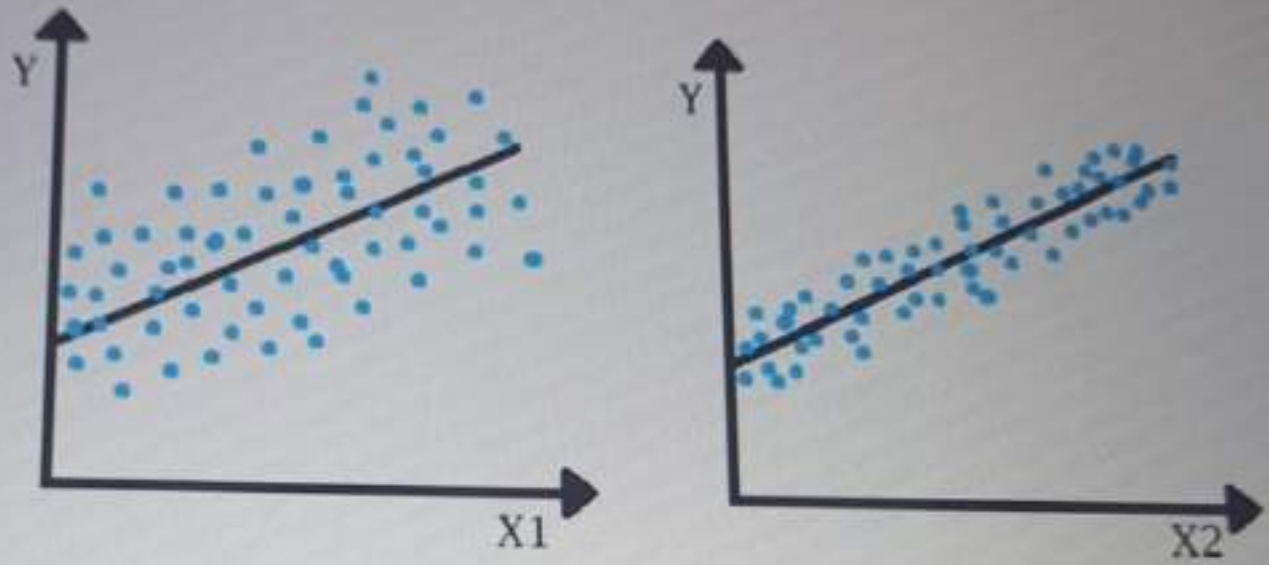
Select any one option

- 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

27. C2 linear Regression

Previous Next

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

- 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

tions

more options

s preferred to MapReduce for processing numerous small files, as it will reduce the overhead in multiple
tions.

ce can be more cost-effective than Spark for an extremely large data set that does not fit in the spa

e is preferred to Spark for iterative processing, as it is much faster than Spark, as it can carry out

rred to MapReduce to create live dashboards, as Spark's processing speed is much faster th

73 min left

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

Submit Test

33. C2 Clustering

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

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

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

Submit Test

36. C2 linear Regression

What does standardised scaling do?

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

ny one option

II, IV, V

I, IV

, IV, V

IV, V

. Which of the following statements is/are correct regarding dataframes?

Media content like images and videos should be processed with unstructured APIs.

When the data schema is not defined, dataframes should be used above RDDs.

Structured APIs have libraries built on top of them to allow writing code more eas

apReduce-style commands in RDDs give better control to analysts over how a partic

d be done.

mes have in-memory processing capabilities as they are built on top of RDDs and, therefore, the

ions

option

2. C3 Spark Slot 1

Next

Which of the following methods can be used to convert a Spark RDD into a Spark DataFrame?

Answer Options

Select any one option

Clear Answer

- RDD.createDF()
- RDD.convertDF()
- ✓ RDD.toDF()
- It is not possible to convert an RDD into a DataFrame as an RDD does not contain a schema, while a DataFrame contains a schema

Answer Options

Select any one option

- Type 1 virtualization
- Type 2 virtualization
- Bare metal virtualization
- Containerization

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, th
are no vaccines or treatments that have been officially approved by WHO after clinical trials. India has not seen the peak of infection ye
and the number of infections is touching a new height daily. The business unit of an Indian health and hygiene company approaches yo
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 t
problem statement above
Statement 2: One of the possible hypotheses for the above problem statement:There is a rise in the number of companies manufacturi
normal/surgical masks due to which the sales of the client's company is decreasing.

Answer Options

Select any one option

Clear An

- 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***

benefit of cloud computing?

Answer Options

Select one or more options

- [x] Rapid elasticity and scalability
- [x] On-demand self-service.
- [x] Resource pooling
- [ ] Access only over a peer-to-peer network connection (A peer-to-peer (P2P) network is created when two or more PC share resources without going through a separate server computer.)

3. C3 Spark Slot 1

< Previous | Next >

While performing word count examples using Spark, Mr Bean wants to split every line on the basis of whitespace and create an RDD of words out of it. What could be the best possible option to achieve the same?

nswer Options

elect any one option

Clear Answer

- Map
- Filter
- FlatMap
- ReduceByKey

```
41)), ("Dhoni", ("M4", 31))])

2. total_score = score.map(lambda x: (x[0], x[1]
[1])).reduceByKey(lambda x, y: x+y)

3. avg_score = total_score._____(lambda x:x/4)

4. avg_score.collect()
```

Answer Options

Select any one option

- map()
- flatMap()
- reduce()
- mapValues()

ne option

hybrid cloud model, maintain the image catalog in a private cloud, and move the web applicati

ybrid cloud model, maintain the web application in a private cloud, and transfer all the imag

thing in a public cloud.

erything in a private cloud.

Suppose your organization has a set of web applications that get a highly varying incoming traffic along with catalog of 80 petabytes, which will be used by these applications. If your organization needs to move to the following would be a cost-effective method to achieve this? (Consider the fact that the public cloud is a che can have privacy issues.)

swer Options

t any one option

Use a hybrid cloud model, maintain the image catalog in a private cloud, and move the we

a hybrid cloud model, maintain the web application in a private cloud, and transfe

50 min left

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

32. C2 Logistic Regression

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

Answer Options

Select any one option

- It helps in finding the different predictive patterns for the different segments that might be present i
- 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

50 min left

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

32. C2 Logistic Regression

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

Answer Options

Select any one option

- It helps in finding the different predictive patterns for the different segments that might be present i
- 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

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

- 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

ark Slot 1

Previous Next

alysing a Spark program and identify that it is taking more than the expected time to execute. The reason for this issue is that it ng some DataFrames repeatedly for processing the other DataFrames. Spark allows you to avoid this by storing DataFrame in o that Spark does not need to recreate it.
tegy would you use here to store the DataFrame in memory?

tions

ne option

Clear Answer

heckpoints to store DataFrame in HDFS

the DataFrame that has been used multiple times

temp tables of DataFrame

all the data frames and combine all the queries in a single DataFrame query

t will be the best data type definition for MYSQL when a field is alphanumeric and has a fixed length?

er Options

any one option

VARCHAR

HAR

ONG

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

35. C3 Advanced ML

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

Answer Options

Select any one option

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 di

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

14. C3 Spark Slot 1

What does the code given below signify in PySpark?

```
lines = sc.textFile('<path to input file, where file actually exists>')
output = lines.map(lambda x:(x.split(" ")[0],x))
```

Answer Options

Select any one option

Clear A

- Splitting the lines of a file based on the space between words and retaining only the first word out of the given line
- Splitting the lines of a file based on the space and retaining all words except the first word out of the given line
- Creating a paired RDD, with the first word as the key and the line as the value
- Creating a paired RDD with the first word as the value and the line as the key

ose your company wants to move its computing infrastructure to cloud but does not want to make a huge upfront investment. Am
llowing models, which one would be the most cost-effective option for your company?

r Options

ny one option

Clear

ommunity cloud model

blic cloud model

vate cloud model

ne of them

Which of the following statements are TRUE about an SQL query?

P: An SQL query can contain a HAVING clause even if it does not have a GROUP BY clause
Q: An SQL query can contain a HAVING clause only if it has a GROUP BY clause
R: All attributes used in the GROUP BY clause must appear in the SELECT clause
S: Not all attributes used in the GROUP BY clause need to appear in the SELECT clause

Answer Options

Select any one option

Clear

- P and R
- P and S
- Q and R
- Q and S

on

Clear

Select one or more options

- [ ]
```
SELECT student_name, year
FROM Student a
RIGHT JOIN Branch b
ON a.branch_id = b.branch_id
WHERE branch_name = 'Electrical Engineering');
```

- [x]
```
SELECT student_name, year
FROM Student a
LEFT JOIN Branch b
ON a.branch_id = b.branch_id
WHERE branch_name = 'Electrical Engineering');
```

- [x]
```
SELECT student_name, year
FROM Student
LEFT JOIN Branch
USING branch_id
WHERE branch_name = 'Electrical Engineering');
```

- [ ]
```
SELECT student_name, year
FROM Student
LEFT JOIN Branch
USING (branch_id)
WHERE branch_name = 'Electrical Engineering');
```

2. Which of the following statements is/are correct regarding dataframes?

I. Media content like images and videos should be processed with unstructured APIs.

II. When the data schema is not defined, dataframes should be used above RDDs.

III. Structured APIs have libraries built on top of them to allow writing code more easily.

IV. MapReduce-style commands in RDDs give better control to analysts over how a particular job should be done.

V. Dataframes have in-memory processing capabilities as they are built on top of RDDs and, therefore, the properties ar

Answer Options

Select any one option

- I, II, IV, V
- I, II, IV
- I, III, IV, V
- I, II, IV, V

Consider an application that must be run on four EC2 instances. Out of the four EC2 instances, two of the EC2 instances execute critical software and need to be run all the time. The third EC2 instance hosts the web server, which gets loaded only when the c accesses it from time to time; customer uptime needs to be maintained at 100%. Finally, the last EC2 instance runs a background collates the logs from time to time.

Which would be a cost-effective combination of instances for this purpose?

Answer Options

Select any one option

- Three on-demand instances and one spot instance
- One reserved instance with a partial upfront payment and three spot instances
- Two reserved instances with an upfront payment and one on-demand instance and one spot instance
- Four on-demand instances

Attempted: 5/6

Revisit Later

following statement holds true for an OLAP system

ed in a normalised form.
ems are used for analytical purpose.
unt of data is stored in OLAP as compared to OLTP system but the query takes
ecute.
restriction on data integrity.

Select an option

- P,Q,S
- P,R,S
- Q,R,S
- Q and S

0/66
Answered

Answer Options

Select any one option

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

normal/surgical masks due to which the sales of the client's com

Statement 2: One of the possible hypotheses for the above p

Statement 1: Understanding the change in customer be

following two statements:

ew question:

ent above

espite the number

179 min left

1. C2 linear Regression

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?

Answer Options

Select any one option

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

...s the tradeoff between the True Positive Rate (TPR) and the False Positive Rate (FPR). The following function is written ...
...using the 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

1/66

Answered

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

Regression

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

- 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

172 min left

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

it Test

< Previous Next >

5. C2 Logistic Regression

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 (β_0) = -9.348 and coefficient of FICO score = 0.0146. Given those parameters, can you calculate the probability of a loan getting approved for someone with a FICO score of 640?

Answer Options

Clear Answer

Select any one option

- 0.35
- 0.40
- 0.45
- 0.50

Answer Options

Select any one option

The cluster centers that are computed in the K-mean

Standardization of the data is important before applying Euc

The centroid of a column with data points 25,

The Euclidean distance betwee

Course 3

1 2 3

15 18 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37

170 min left

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

Submit Test

8. C2 linear Regression

Previous Next

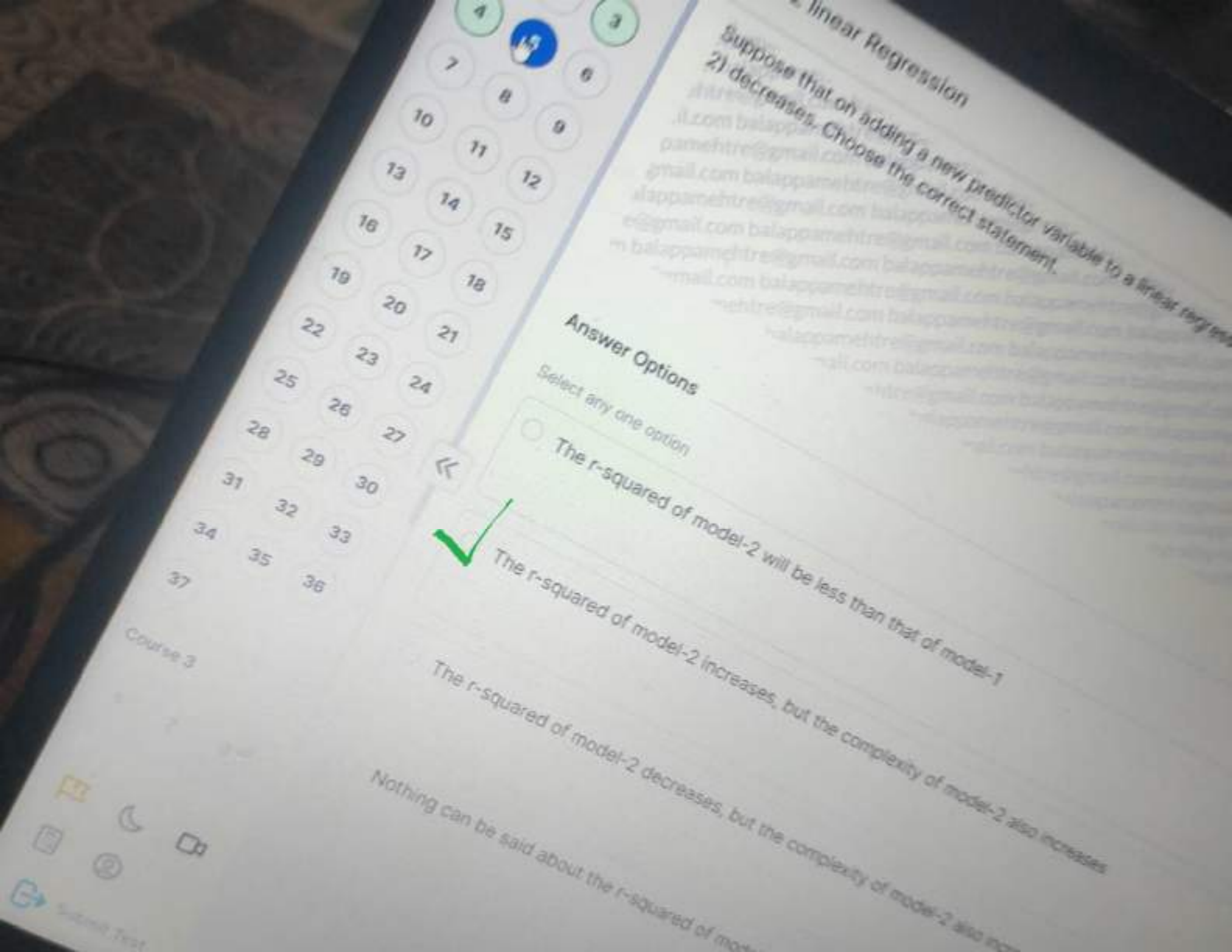
Suppose that on adding a new predictor variable to a linear regression model (model-1), the adjusted r-squared of the new model (model-2) decreases. Choose the correct statement.

Answer Options

Select any one option

Clear Answer

- 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


Linear Regression
Suppose that on adding a new predictor variable to a linear reg
2) decreases. Choose the correct statement.
Answer Options
Select any one option
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
Nothing can be said about the r-squared of
Course 3

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 interpretation is not possible for K-Means clustering

Submit Test

has not seen the peak of infection yet and the number of
an Indian health and hygiene company approaches you to
number of corona infections increasing daily".

Answer the below question:
Consider the following two statements:
Statement 1: Understanding the change in customer beh
understanding for the problem statement above
Statement 2: One of the possible hypotheses for the ab
companies manufacturing normal/surgical masks due to

2/74

nswered

Answer Options

Select any one option

Statement 1 is correct and Statement 2 is wrong

Statement 2 is correct and Statement 1 is wrong

Both the statements are correct

Answer Options

Select any one option

2/74

Answered

- b(i) is the average distance from the nearest ne
- a(i) is the average distance from own cluster(
- If S(i) = 1 then the datapoint is similar to its ow
- ✓ Silhouette Metric ranges from 0 to +1

Submit

Answer Options

Select any one option

2/74

nswered

- The probability of the customer not churning is 3
- The probability of the customer churning is 3 tim
- The probability of the customer not churning is
- The probability of the customer churning is 4 t

Submit

Some of the independent variables (predictors) i
independent variable in the model is redundant.
Suppose that you are building a multiple linear re
statements is TRUE w.r.t. multicollinearity?

Answer Options

Select any one option

Multicollinearity is a problem when your ind
variables

Multicollinearity is a problem when your go
riables

Multicollinearity is not a problem if a variab

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

Clear Answer

Select any one option

```
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)
```

Answer Options

Select any one option

0

0.25

0.5

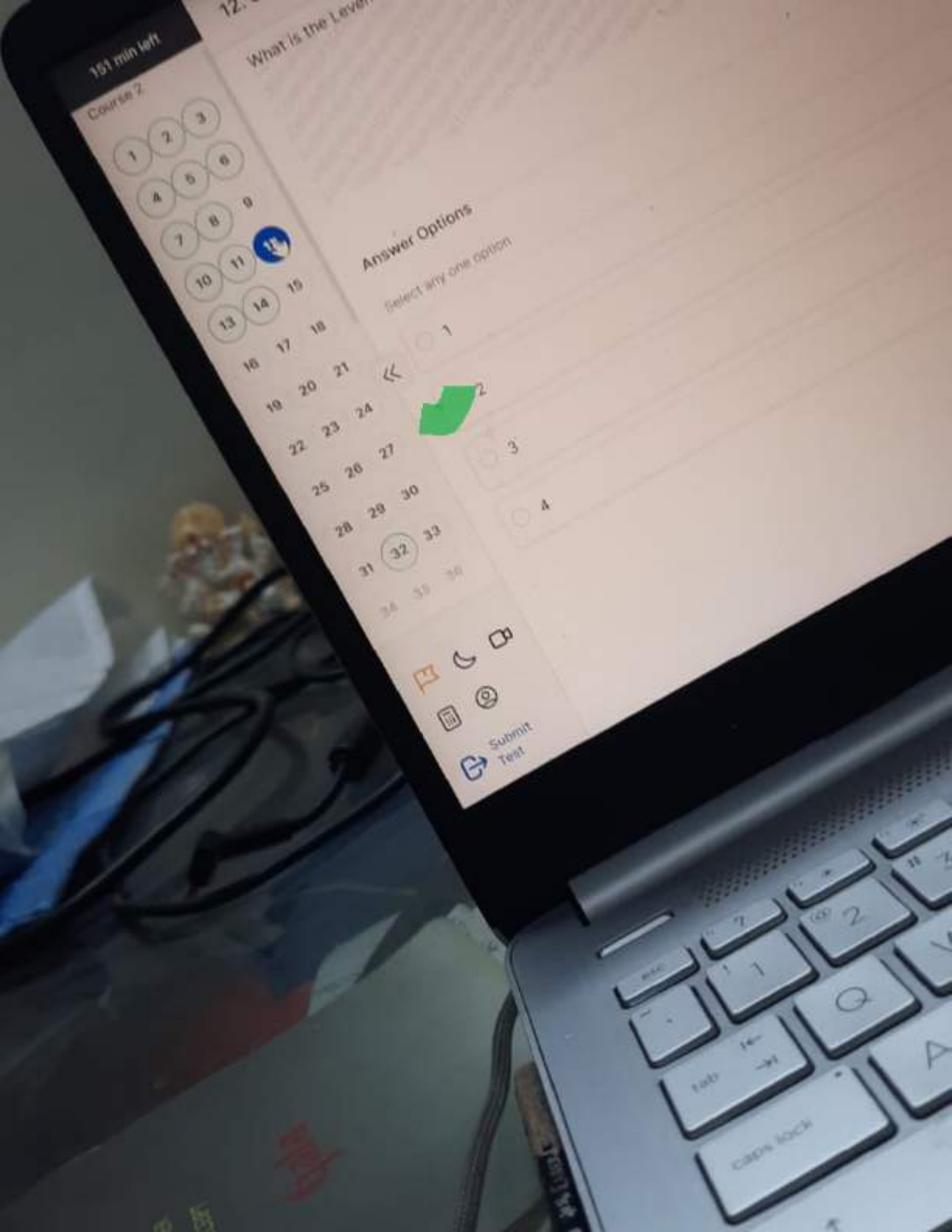
1

15 18 21 24 27 30 33 36

22 25 28 31 34 37

Course 3

2 3


151 min left
Course 2
What is the Level
Answer Options
Select any one option
1
2
3
4
Submit Test

and the number of infections is touching a new [illegible]
know! Why the sales of masks is decreasing despite the number of corona infections [illegible]
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

Clear Ans

Select any one option

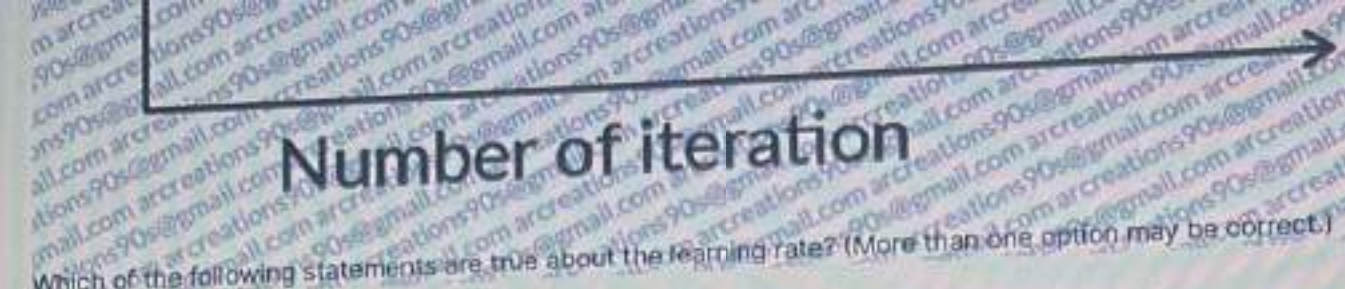
- Neural Network
- Logistic regression
- Decision tree
- All of the above

mit Test

Answer Options

Select any one option

- A
- B
- C
- D

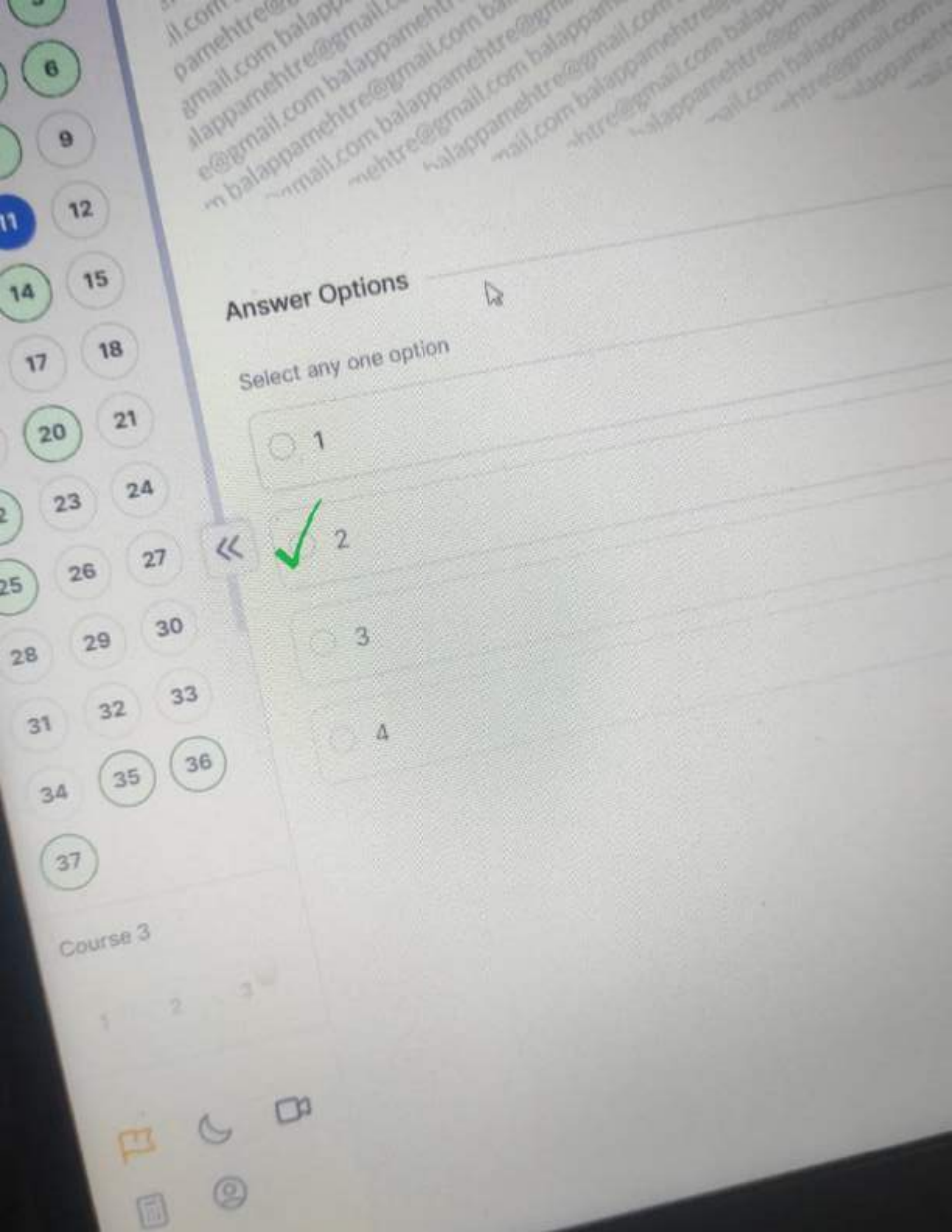


Which of the following statements are true about the learning rate? (More than one option may be correct.)

Answer Options

Select one or more options

- [x] The learning rate of curve C is highest among all curves
- [ ] The learning rate for curve B is lower than A
- [x] The learning rate for curve B is higher than A
- [ ] The learning rate of curve C is the smallest among all curves


Answer Options
Select any one option
1
2
3
4
Course 3

Answer Options

Select any one option

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

VIF (Variance Inflation Factor) is used to

Answer Options

Select any one option

- 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 pr
- If a variable is a product of two other variables, it can have a

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 understa problem statement above.

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

Answer Options

Select any one option

- 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

Test

For a completely random binary classification model, what wi

Answer Options

Select any one option

- 0
- 0.25
- 0.5
- 1

31/74
Answered

Submit

The difference between '+' and '*' quantifiers is _____

Answer Options

Select any one option

'+' 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

Submit Test

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

Answer Options

Select any one option

- '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 dendogram
- All the above

Answer Options

Select one or more options

Clear A

- [ ] 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.

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 Answer

- A, B, C and D
- A, C, and D
- A, B and C
- B, C and D

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

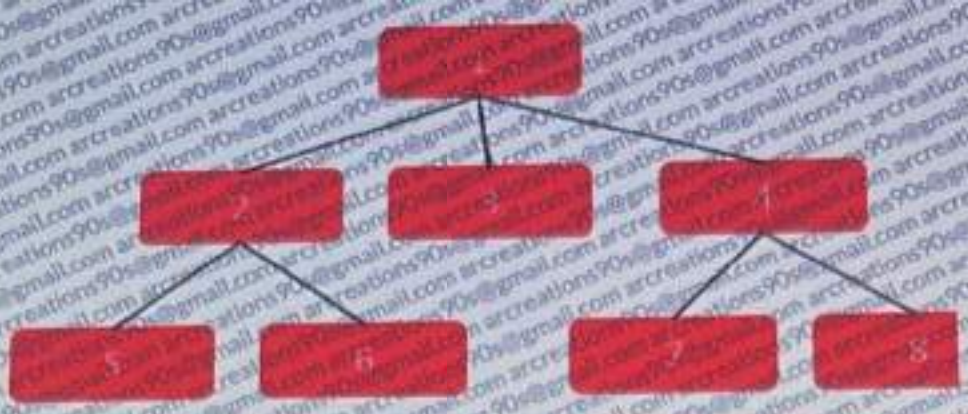
Which among the following is the highest for the given confusion matrix?

Answer Options

Select any one option

- Sensitivity
- Specificity
- Precision
- Accuracy

Submit Test



I. Node 8 is the root node
II. The number of leaf nodes is 5
III. Nodes 2, 3 and 4 are internal node
IV. Nodes 2 and 4 are internal node

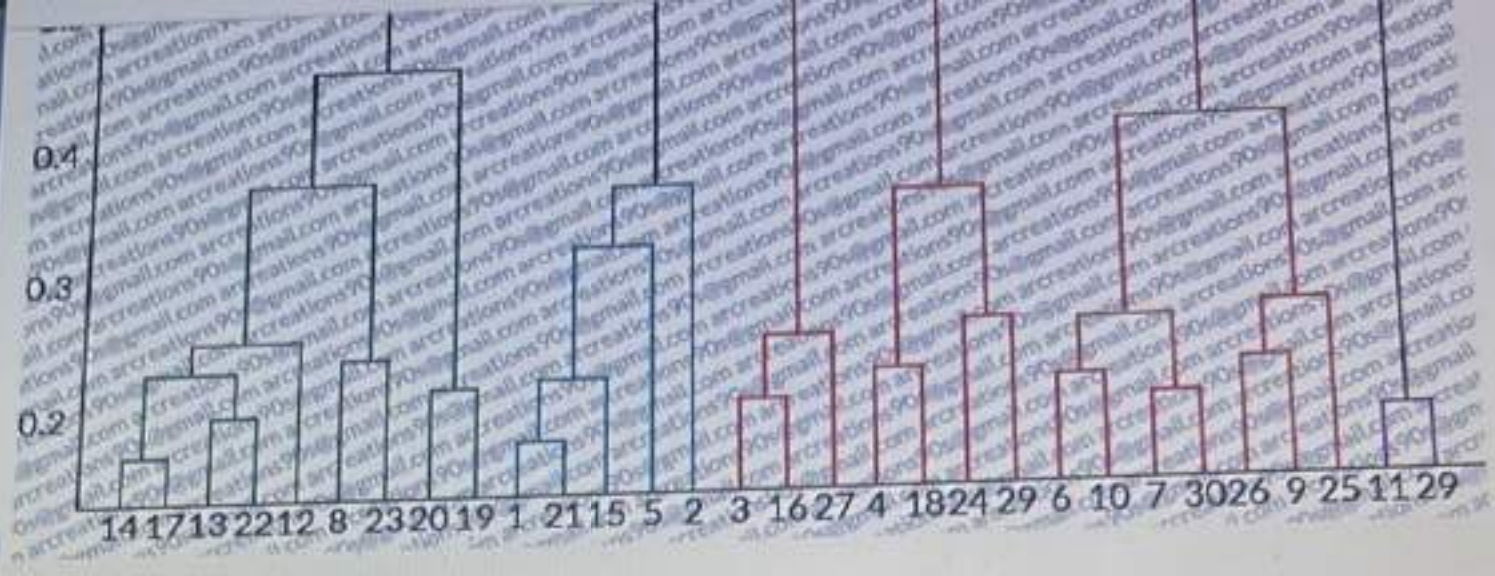
II & IV

Answer Options

Select any one option

- Only I
- Only II

Submit Test



Answer Options

Clear Answer

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 interpretation is not possible for K-Means clustering

Answer Options

Select any one option

- 65,536
- 1,024
- ✓ No limit
- 9,999

Answer Options

Select any one option

- Type 1 virtualization
- Type 2 virtualization
- Bare metal virtualization
- Containerization ✓

Answer Options

Select any one option

- Convert the CSV file to a JSON file to reduce the file size
- Apply the gzip compression technique on the CSV file
- Convert the CSV to parquet format with snappy compression on it
- CSV files cannot be reduced further by applying any compression technique

Consider an application that must be run on four EC2 instances. Out of the four EC2 instances, two of the EC2 instances execute mission-critical software and need to be run all the time. The third EC2 instance hosts the web server, which gets loaded only when the customer accesses it from time to time. Customer uptime needs to be maintained at 100%. Finally, the last EC2 instance runs a background job that collates the logs from time to time.

Which would be a cost-effective combination of instances for this purpose?

Answer Options

Select any one option

Clear Answ

- Three on-demand instances and one spot instance
- One reserved instance with a partial upfront payment and three spot instances
- Two reserved instances with an upfront payment and one on-demand instance and one spot instance
- Four on-demand instances

Answer Options

Select any one option

- 2 and 3
- 3
- 2
- 1,2 and 3

wants to implement an aggressive customer retention campaign to retain th
launched extremely low-cost mobile plans, and you want to avoid churn as
budget is not a constraint.

Problem statement 2: Let's say you are building a cancer detection model w
patient who has not cancer can be detected correctly. It can have serious in
wrongly detected as "not cancer", the patient will die of cancer, and if wrong

Problem statement 3: You have to build an image classification model where
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 classif

Answer Options

Select any one option

Problem statement 1: Specificity

Problem statement 2: Sensitivity

Problem statement 2: Specificity

Problem statement 3: Accuracy

orming segregation or pre-processing of a database before running a logistic regression is important for several reasons:

Data cleaning: Segregation helps in cleaning the database and removing any missing values or outliers that could affect the accuracy of the logistic regression model.
Feature selection: Segregation helps to identify and select the most relevant features that have the most significant impact on the outcome variable. This helps to improve the accuracy of the model and reduce the risk of overfitting.
Data normalization: Segregation can also be used to normalize the data by scaling the features to the same range. This helps to prevent any biases caused by differences in the scales of the features.
Balance of classes: Segregation can be used to ensure that the classes are balanced, especially when dealing with imbalanced datasets. This can help to prevent the model from being biased towards the majority class.

Overall, performing segregation or pre-processing of a database before running a logistic regression on it helps to improve the quality of the data and ensure that the model is accurate and reliable.

IV option

Regenerate response

	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

- Accuracy
- Precision
- Sensitivity

What is the use of performing segmentation on a dataset before running a logistic regression on it?

Answer Options

Select any one option

- 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 migh
- It helps capture the trends easily when there is a class imbalance

	Actual Positive	Actual Negative
Predicted Positive	196	28
Predicted Negative	23	254

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

Answer Options

Select any one option

Sensitivity

Specificity

Precision

Accuracy

Some of the independent variables (predictors) might be interrelated, due to which the presence of a particular independent variable i
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 T
w.r.t. multicollinearity?

Answer Options

Select any one option

Clear An

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 m
re categories

options

oothesis for a simple linear regression model is H0: $\beta 1 = 0$

e turns out to be greater than 0.05 for $\beta 1$, it means $\beta 1$ is significant

to be insignificant, that means there is no relationship between the dependent and inde

urns out to be less than 0.05 for $\beta 0$, it means that $\beta 0$ is non-zero

istic is used to determine

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

ple correct answer

ar regression model when you fit a straight line through the data you'll get the two parameters
t β0 and the slope β1. Which of the following is true for β0 and β1? (More than one option may

s for a simple linear regression model is H0: β1 = 0

ut to be greater than 0.05 for β1, it means β1 is significant

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

3

6

9

12

15

18

21

24

27

30

33

36

Answer Options

Select any one option

Entropy

Information Gain

Gini Index

None of these

Submit Test

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

Submit Test

Multiple options correct

Previous | Next

[illegible]

```
[illegible]
```

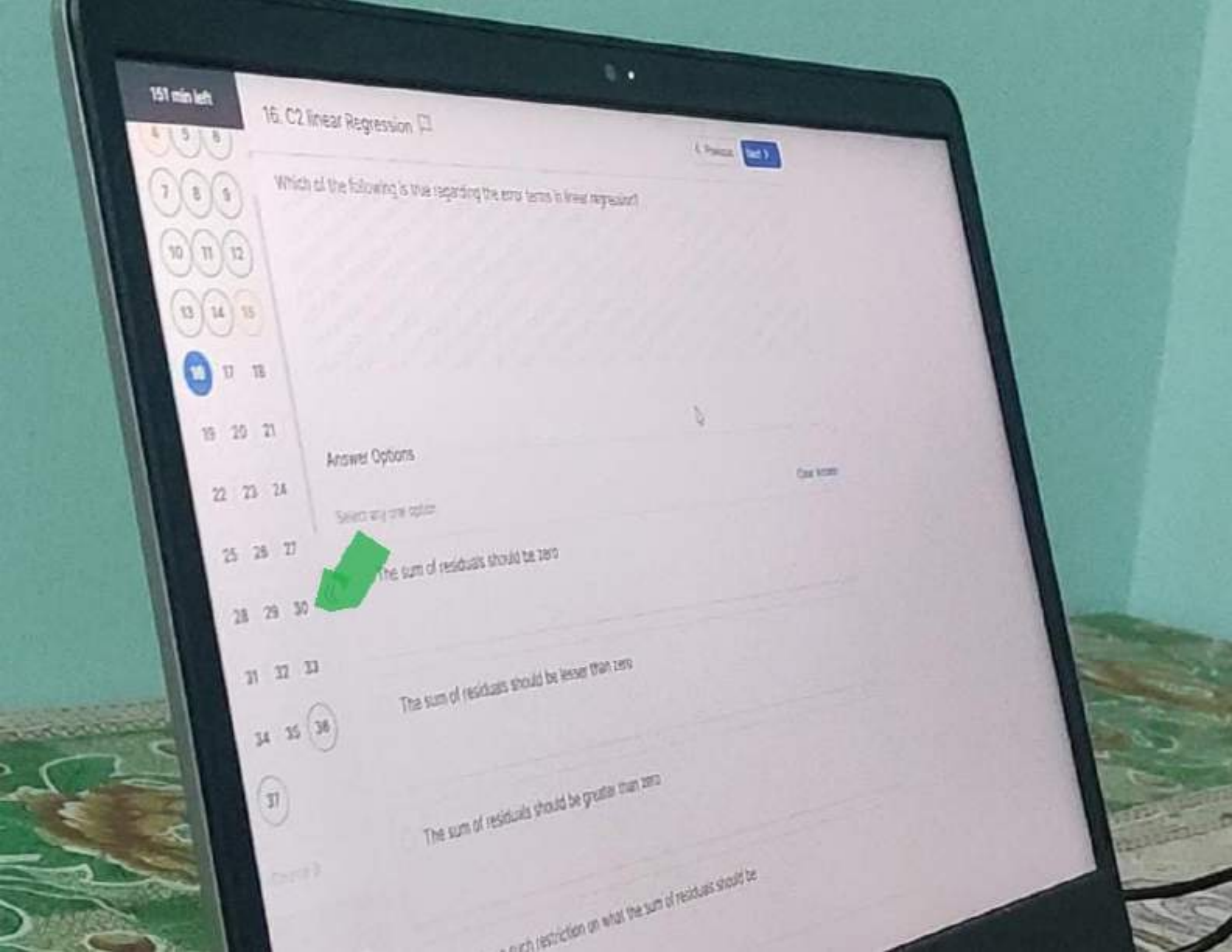
Which of the following statements are true? (More than one option may be correct)

Answer Options

Select one or more options

Clear Answer

- [ ] 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


151 min left
16. C2 linear Regression
Which of the following is true regarding the error terms in linear regression?
Answer Options
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

8. C2 linear Regression

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

Answer Options

Select any one option

- The relationship between the two variables is not strong
- The correlation coefficient between the two variables is also 0.47
- 47% of the variance in the independent variable is explained by the dependent variable
- 47% of the variance in the dependent variable is explained by the independent variable.

50 min left

15. C2 linear Regression

Previous Next

How is regression different from classification?

Answer Options

Select any one option

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

5 6 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

What does standardised scaling do?

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

Submit Test

143 min left

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

9. C2 Logistic Regression

< Previous Next >

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

Answer Options

Select any one option

Clear Answer

- 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

141 min left

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

Submit Test

10. C2-Basics of NLP and Text Mining

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

Answer Options

Select any one option

'+' 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

What is the Levenshtein distance between 'decade' and 'dictate'?

Answer Options

Select any one option

3

4

5

6

11. C2 linear Regression

Previous

VIF (Variance Inflation Factor) is used to detect Multicollinearity. Which of the following statements is NOT true for VIF?

Answer Options

Select any one option

- 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

138 min left

12. C2 Multiple options correct

Previous Next

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 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

Clear Answer

Select one or more options

- 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

Submit Test

Answer Options

Select any one option

- 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

6. C2 Multiple correct answer

Previous

Answer Options

Select one or more options

```
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()
```

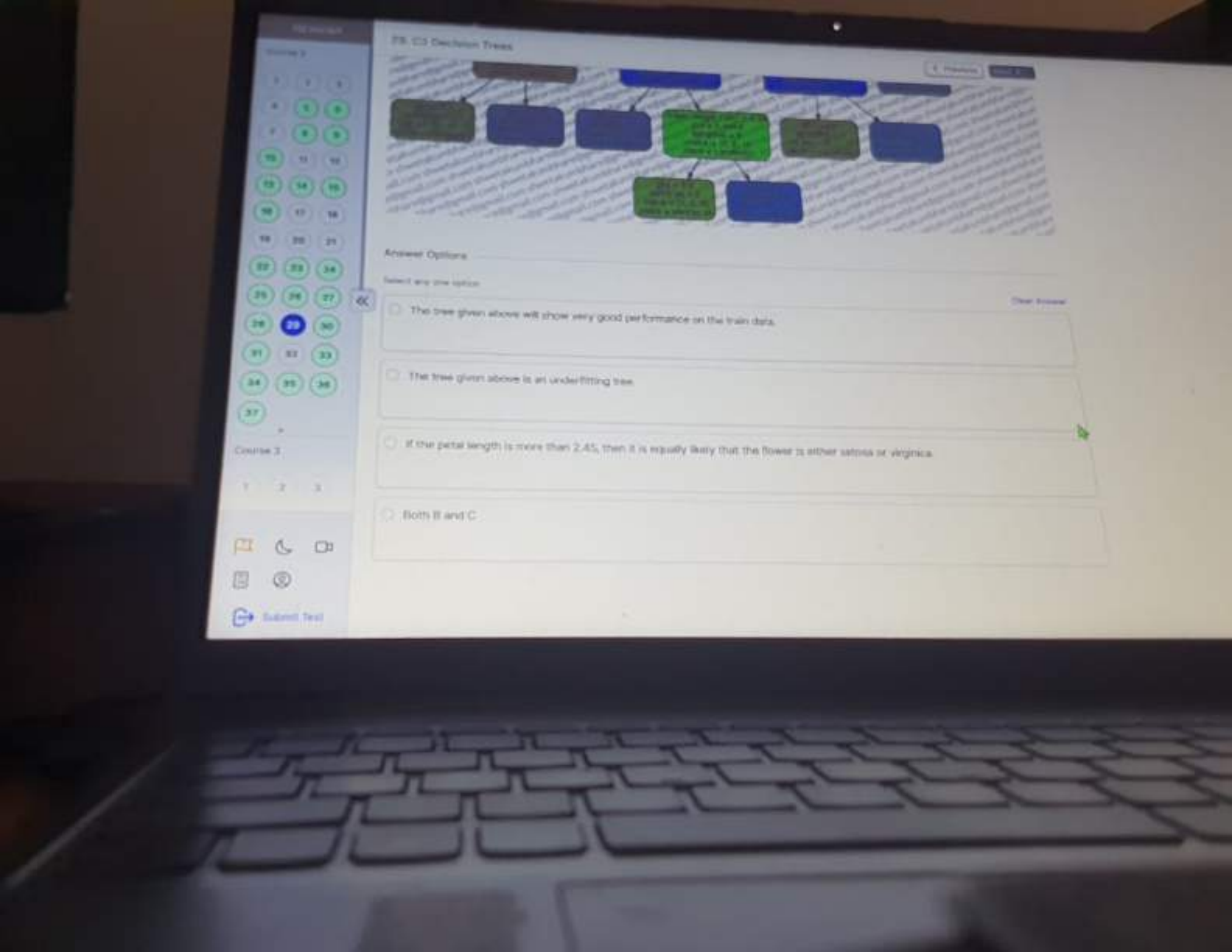
Submit Test

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 (B0) = -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?

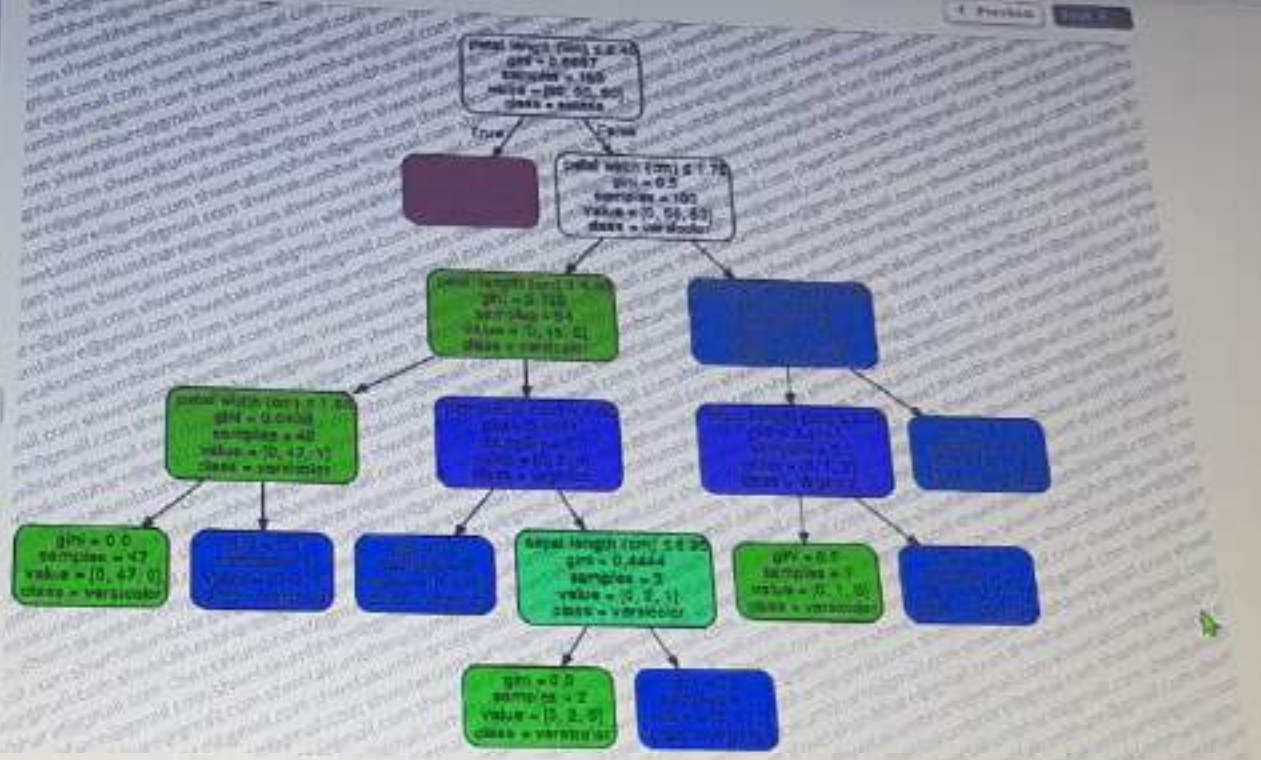
Answer Options

Select any one option

- 0.35
- 0.45
- 0.55
- 0.65


29. C3 Decision Trees
Answer Options
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
Submit Test

29. C2 Decision Trees



Answer Options

Select any one option

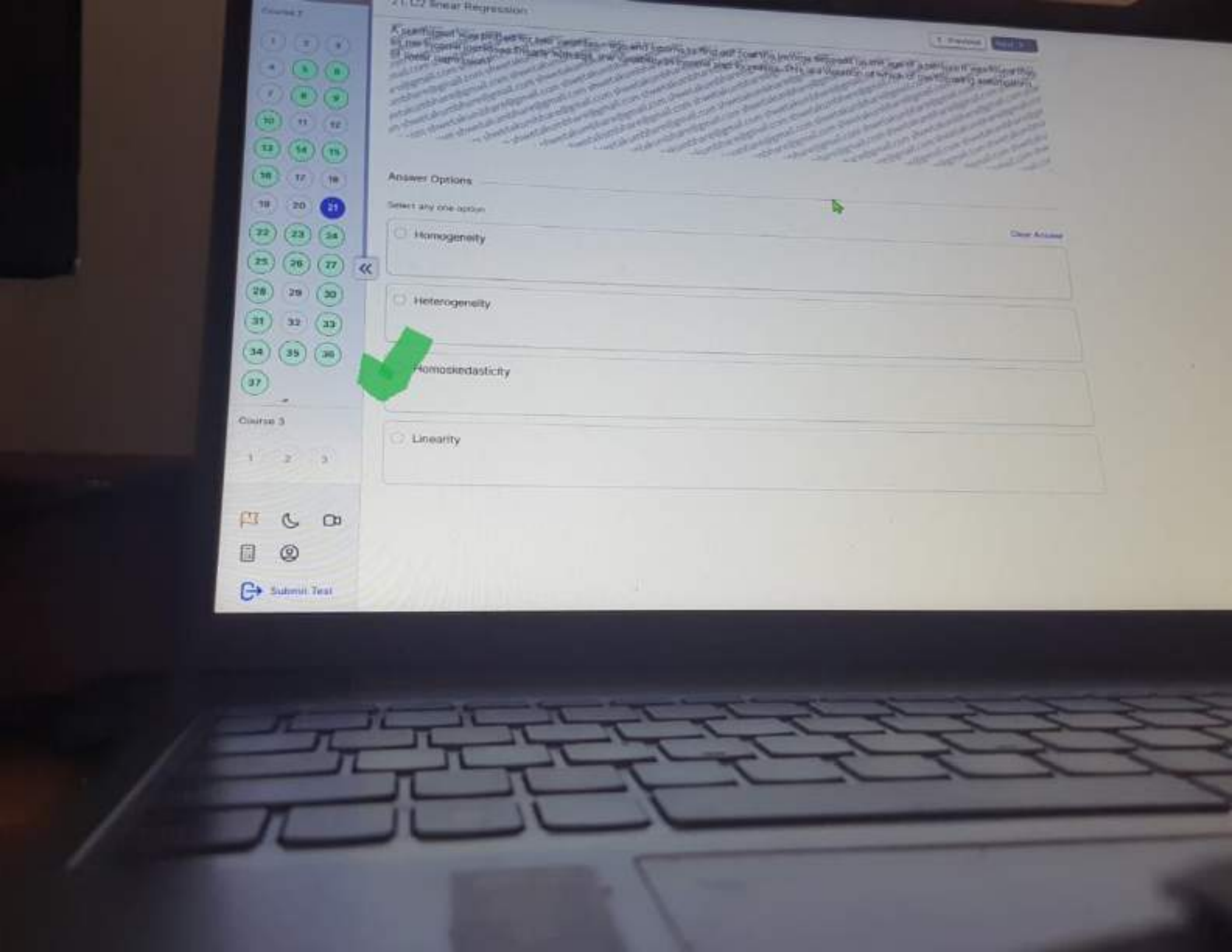
Clear Answer

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

Course 2

Course 3

Submit Test


Answer Options
Select any one option
Homogeneity
Heterogeneity
Homoskedasticity
Linearity
Submit Test

20. C2 linear Regression

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

- 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

Course 3

Submit Test

10. C2 Multiple correct answer

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)

Previous Next

Answer Options

Select one or more options

Dendrogram inspection method

Clear Answer

Elbow Method

Single Linkage Method

Silhouette score

Answer Saved

wer Options

t any one option

Hopkins statistic decides if the data is suitable for clustering or not

Hopkins statistic lie between -1 and 1

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

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

more options

ogram inspection method

Method

nkage Method

e score

124 min left

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

Submit Test

16. C2 Logistic Regression

Previous Next

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

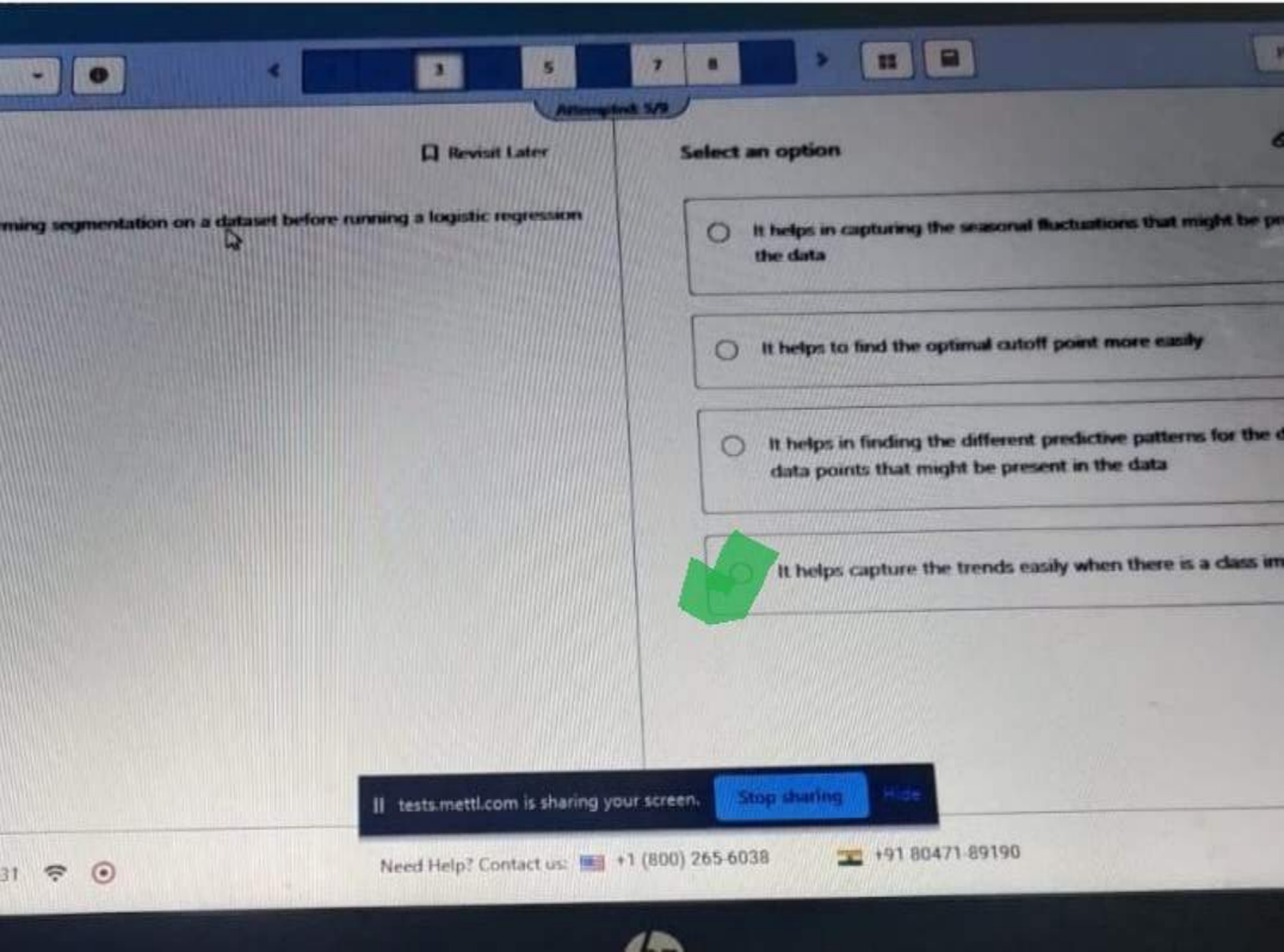
Clear Answer

0

0.25

0.5

1


Revisit Later
Select an option
rming segmentation on a dataset before running a logistic regression
It helps in capturing the seasonal fluctuations that might be pr
the data
It helps to find the optimal cutoff point more easily
It helps in finding the different predictive patterns for the
data points that might be present in the data
It helps capture the trends easily when there is a class im
tests.mettl.com is sharing your screen.
Stop sharing
Hide
Need Help? Contact us:
+1 (800) 265-6038
+91 80471 89190

Question 7

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.

Find the values of A and B using the MA process given SIGMA = 0.5 and MU = 15.
(Note: Multiple options might be correct)

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

Answer Options

Select one or more options

- [ ] B = -1.25
- [x] A = 15.5
- [x] B = 1.75

17. C2 Clustering

Which of the following statements is NOT true?

Answer Options

Select any one option

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

18 C2 Clustering

Previous Next

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 Answer

Single linkage and Complete linkage

Complete linkage and Single linkage

Single linkage and Average linkage

Complete linkage and Average linkage

19. C2 Multiple options correct

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

Answer Options

Select one or more options

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

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

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

er Options

any one option

0

0.5

1

100

: Multiple correct answer

< Previous Next >

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

Options

Clear Answer

e or more options

ndrogram inspection method

ow Method

gle Linkage Method

ouette score

142 min left

Course 2

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

14. C2 Clustering

< Previous Next >

Which of the following statements is NOT true?

Answer Options

Select any one option

Clear An

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

35. C2 Logistic Regression

Consider the following univariate logistic model:
$Y = \beta 0 + \beta 1 X1$

Which of the following statements is NOT true?

Answer Options

Select any one option

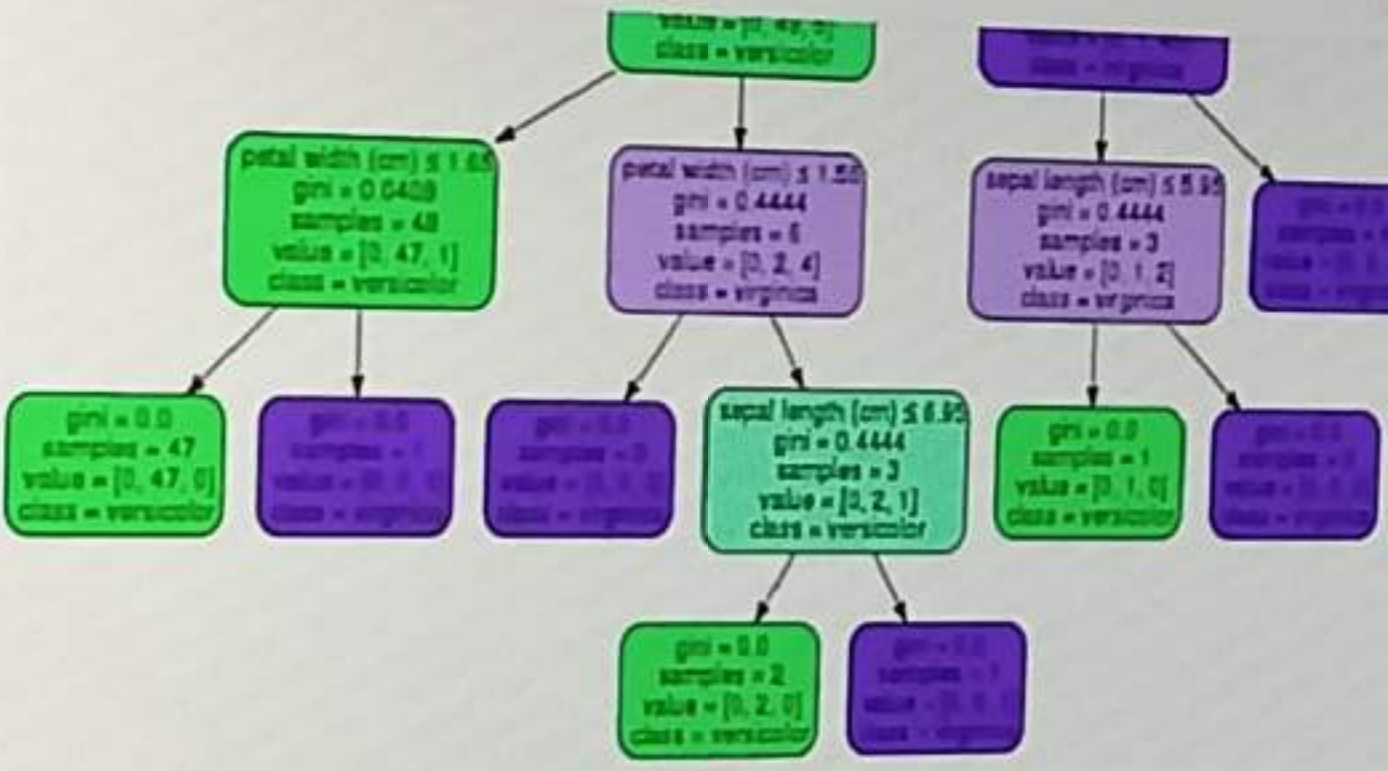
The maximum likelihood estimation determines the best combination of β0 and β1.

If β1 is increased by 1 unit, Y increases by 1 unit.

β0 is the y-intercept

If β1 is increased by 1 unit, log-odds increases by 1 unit.

21. C2-Decision Trees



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.

Both B and C

Submit Test

36. C2 Logistic Regression

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

Answer Options

Select any one option

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.

The learning rate of curve C is highest among all curves

he learning rate for curve B is lower than A

e learning rate for curve B is higher than A

e learning rate of curve C is the smallest among all curves

e of the above.

22. C2 Multiple options correct

Answer Options

Select one or more options

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

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

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

23. C2 Business Problem Solving

he 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. Your will build _____ model as your initial solution.

Answer Options

Select any one option

Clear Answer

- Neural Network
- ✓ Logistic regression
- Decision tree
- All of the above

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

Answer Options

Select any one option

- Entropy
- Information Gain
- Gini Index
- None of these

min left

2 3
5 6
8 9
11 12
14 15
17 18
20 21
23 24
26 27
29 30
32 33
35 36

Submit Test

24. C2 Logistic Regression

Previous Next

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, 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

Clear Answer

Select any one option

Problem statement 1: Specificity

Problem statement 2: Sensitivity

Problem statement 2: Specificity

Problem statement 3: Accuracy

25. C2 Clustering

< Previous Next >

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?

Answer Options

Clear Answer

Select any one option

- '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 dendogram
- ✓ All the above

What does standardised scaling do?

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