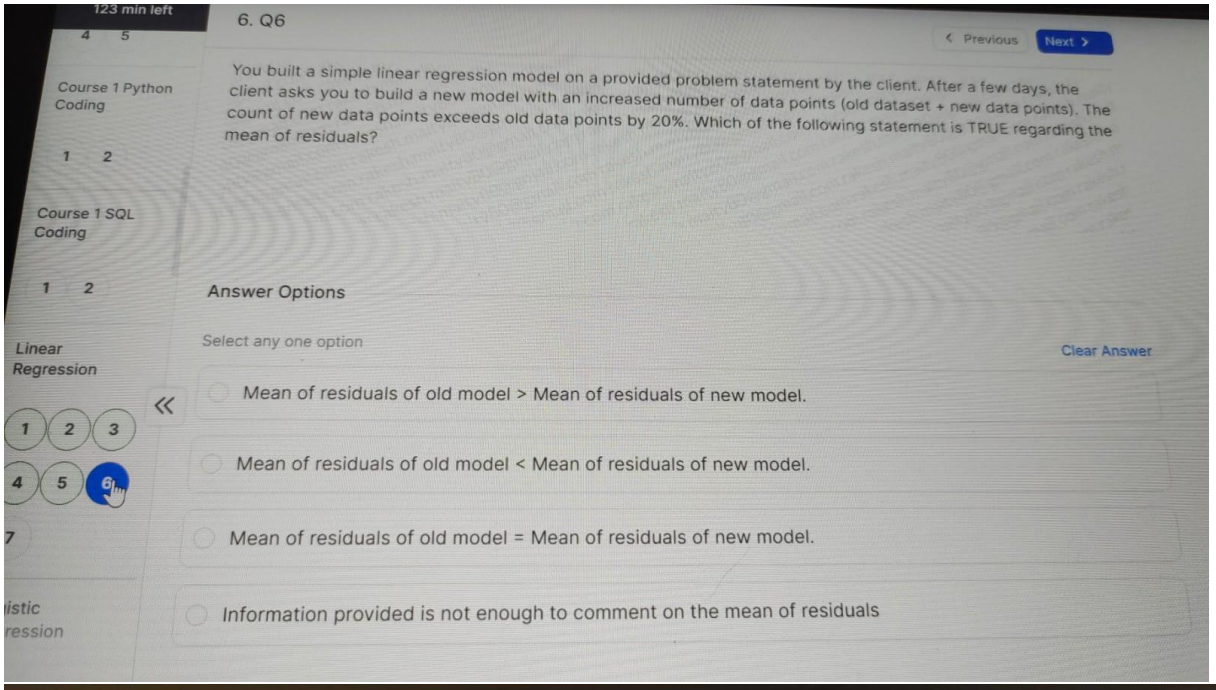
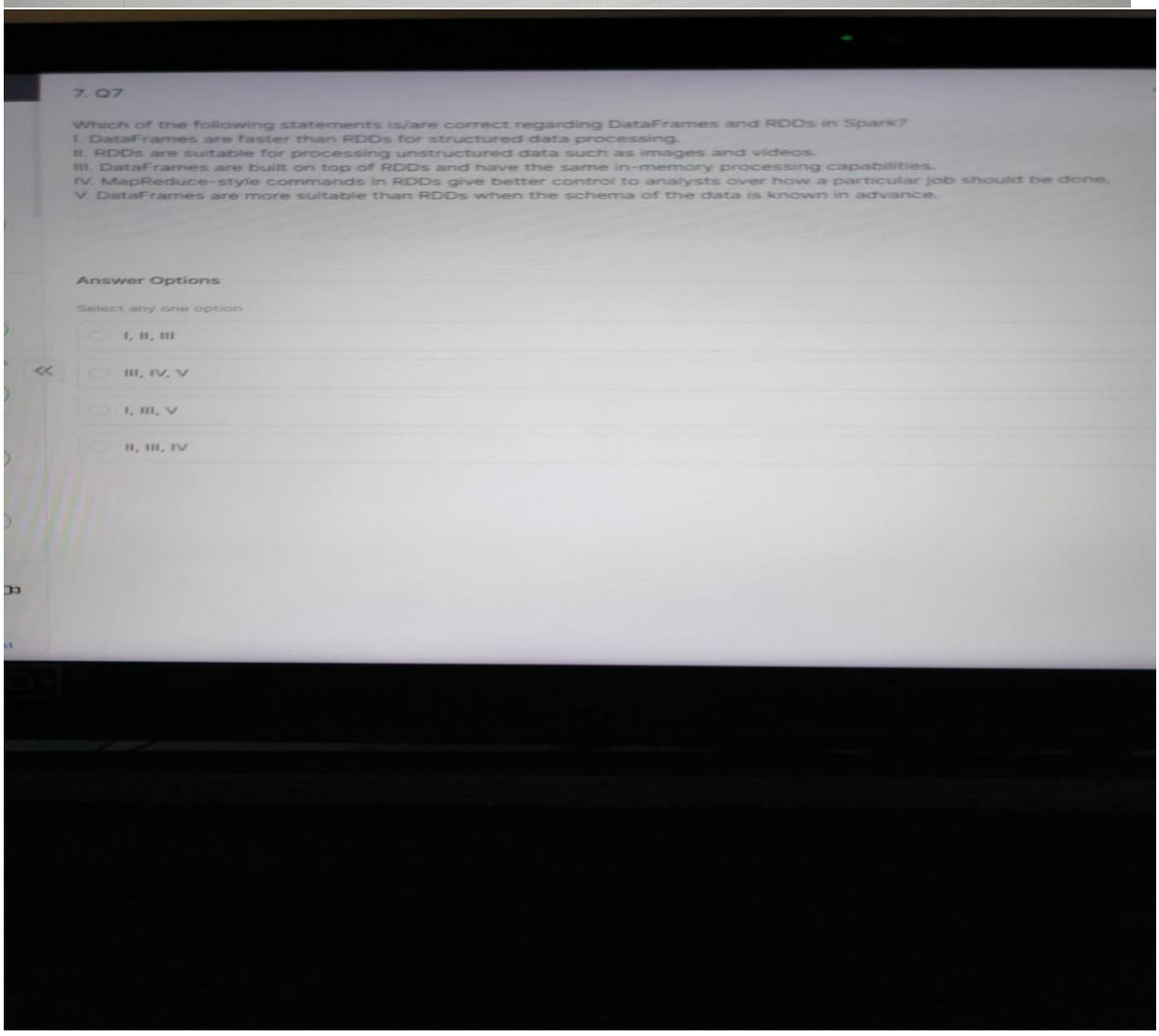

123 min left
4 5
Course 1 Python Coding
1 2
Course 1 SQL Coding
1 2
Linear Regression
1 2 3
4 5 6
7
istic
ression
6. Q6
Previous
Next
You built a simple linear regression model on a provided problem statement by the client. After a few days, the client asks you to build a new model with an increased 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
Clear Answer
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


7. Q7
Which of the following statements is/are correct regarding DataFrames and RDDs in Spark?
I. DataFrames are faster than RDDs for structured data processing.
II. RDDs are suitable for processing unstructured data such as images and videos.
III. DataFrames are built on top of RDDs and have the same in-memory processing capabilities.
IV. MapReduce-style commands in RDDs give better control to analysts over how a particular job should be done.
V. DataFrames are more suitable than RDDs when the schema of the data is known in advance.
Answer Options
Select any one option
I, II, III
III, IV, V
I, III, V
II, III, IV

15. Inferential Statistics

In a particular game series, the following cumulative probability table has bee
particular player in one game.

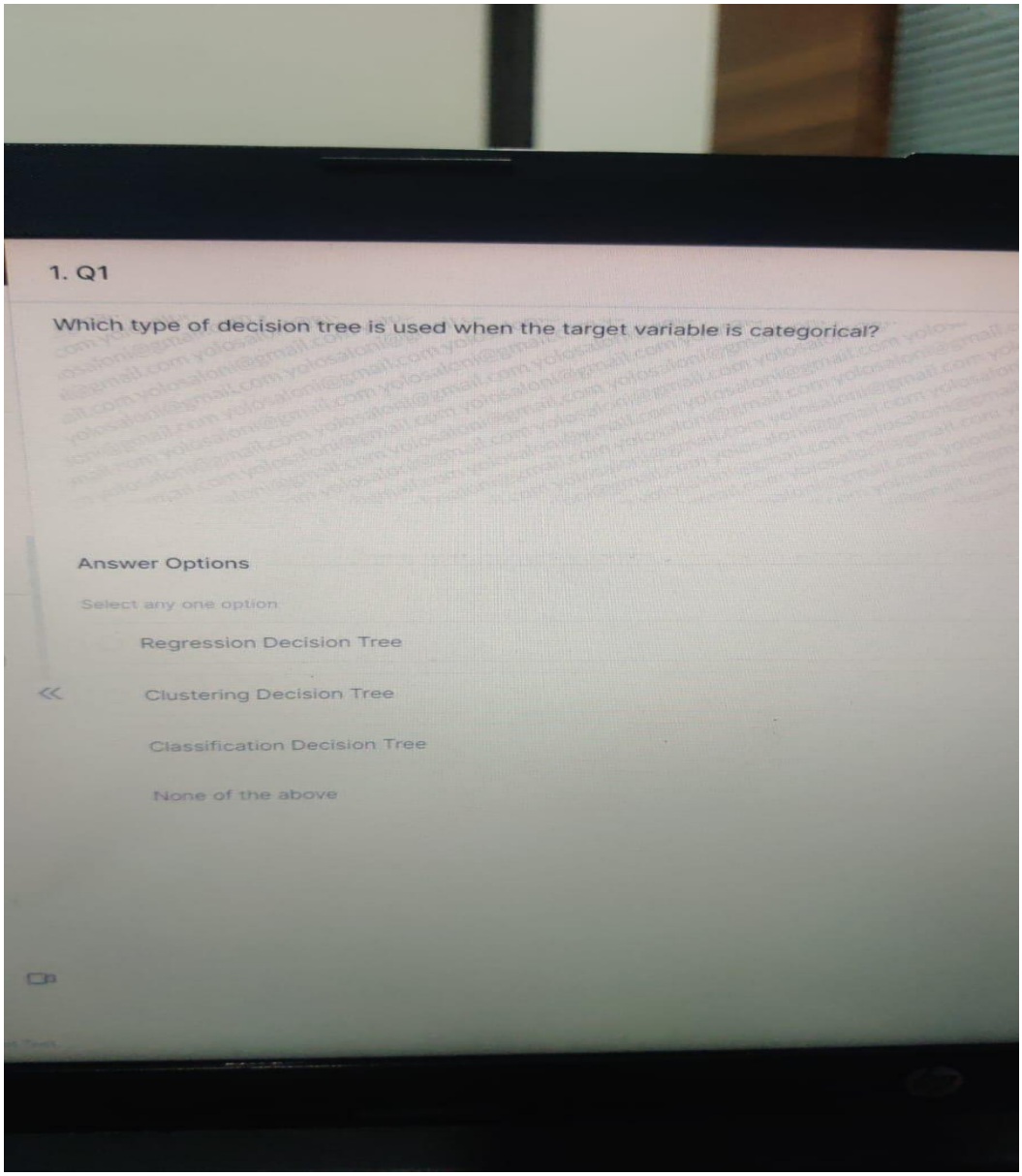
x	F(x) = P(X<=x)
20	0.2
40	0.26
60	0.4
80	0.8
120	0.95
200	1

If the player played 40 games in that series, calculate the number of games in
to 120 points

Answer Options

Select any one option

- 10
- 12
- 14
- 6

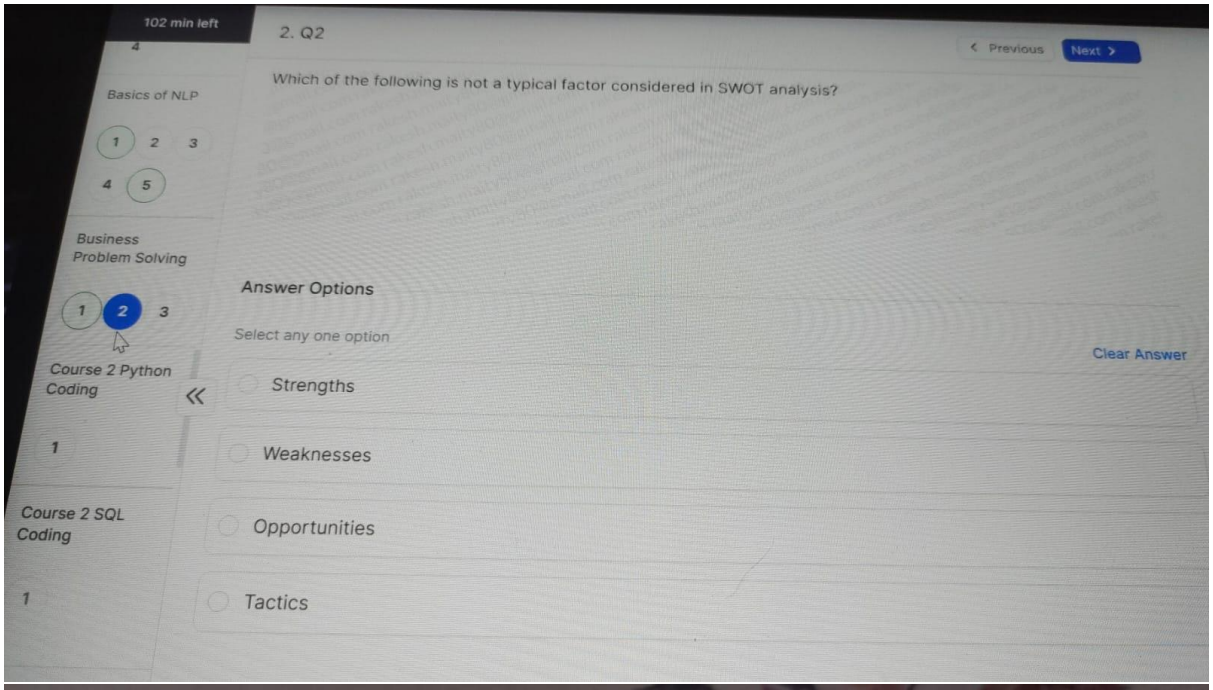

1. Q1
Which type of decision tree is used when the target variable is categorical?
Answer Options
Select any one option
Regression Decision Tree
Clustering Decision Tree
Classification Decision Tree
None of the above

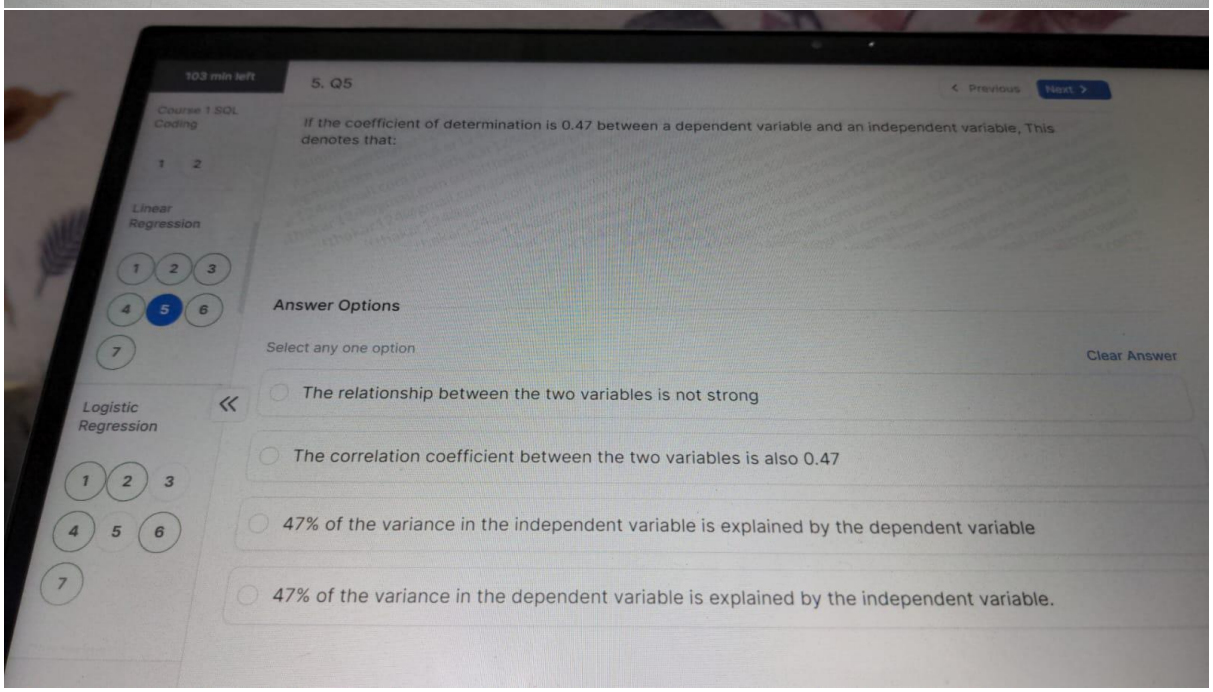
16. Inferential Statistics

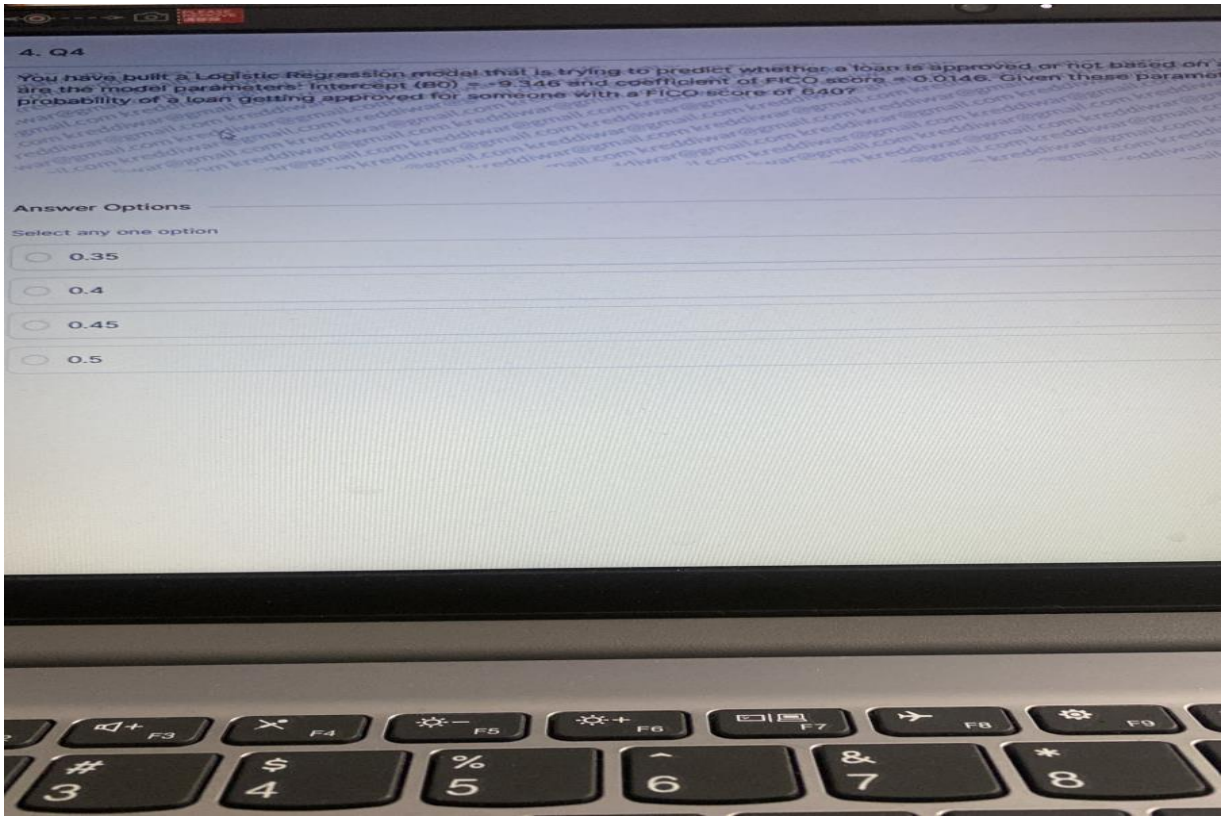
For a random variable that is normally distributed the mean comes out to b
experiment, what would be the probability that the value of this random var
below.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.0
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.03
−1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.04
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.05
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.06
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.08
−1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.09
−1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.11
−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.13
−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.16
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.186
−0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.214
−0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.245
−0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.277
−0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.312
−0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.348
−0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
−0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
−0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

z	.00	.01	.02	.03	.04	.05	.06	.07
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808


102 min left
2. Q2
Previous
Next
Which of the following is not a typical factor considered in SWOT analysis?
Basics of NLP
1 2 3
4 5
Business Problem Solving
1 2 3
Course 2 Python Coding
1
Course 2 SQL Coding
1
Answer Options
Select any one option
Clear Answer
Strengths
Weaknesses
Opportunities
Tactics


103 min left
5. Q5
Previous
Next
Course 1 SQL Coding
1 2
If the coefficient of determination is 0.47 between a dependent variable and an independent variable, This denotes that:
Linear Regression
1 2 3
4 5 6
7
Logistic Regression
1 2 3
4 5 6
7
Answer Options
Select any one option
Clear Answer
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.


4. Q4
You have built a Logistic Regression modal that is trying to predict whether a loan is approved or not based on
are the model parameters: Intercept (B0) = -9.346 and coefficient of FICO score = 0.0146. Given these parame
probability of a loan getting approved for someone with a FICO score of 640?
Answer Options
Select any one option
0.35
0.4
0.45
0.5

1. Q1

Consider the following confusion matrix.

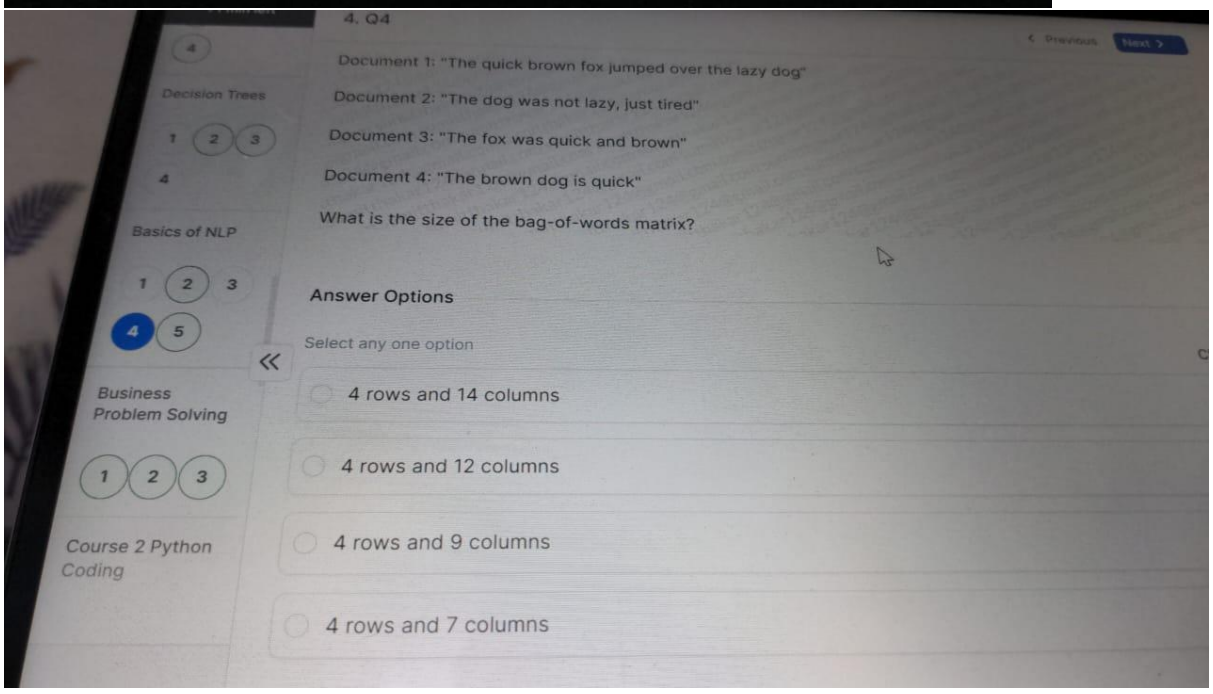
Which among the following is the lowest for the given confusion matrix?

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

Answer Options

Select any one option

- Accuracy
- Precision
- Sensitivity
- Specificity



Previous Next >

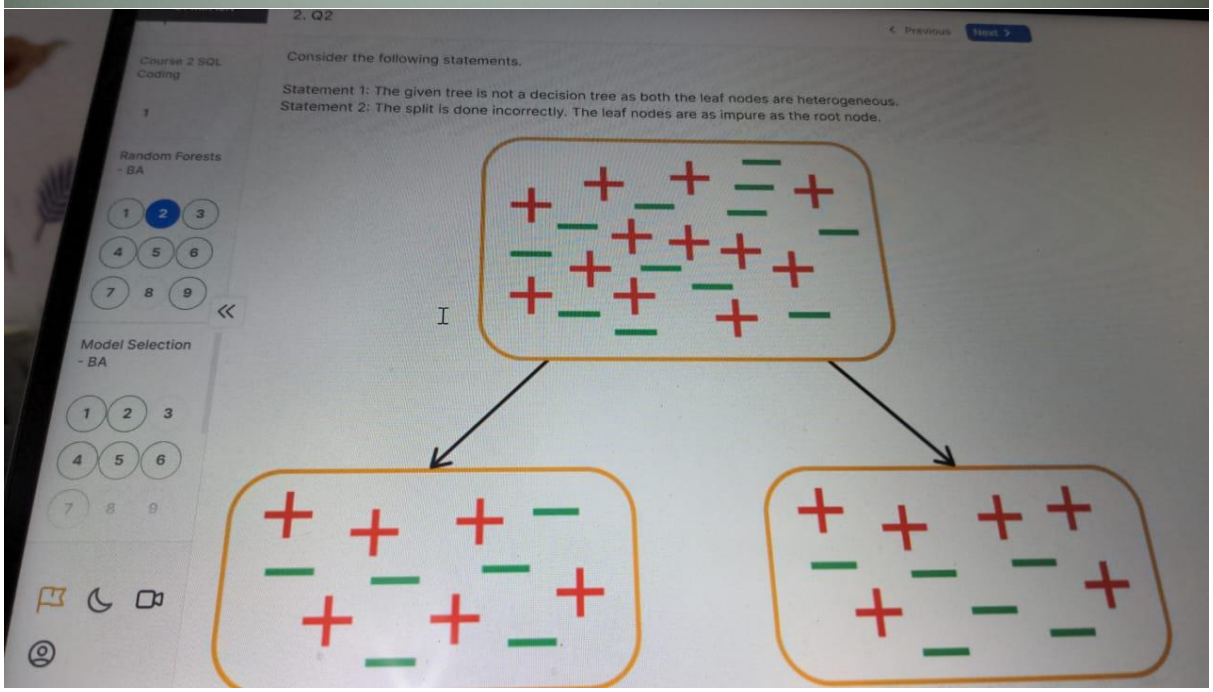
If you are building a model to detect if a person has a specific disease based on their vitals, which of the following algorithms would help the medical practitioner the most? (Medical practitioners prefer models with high interpretability)

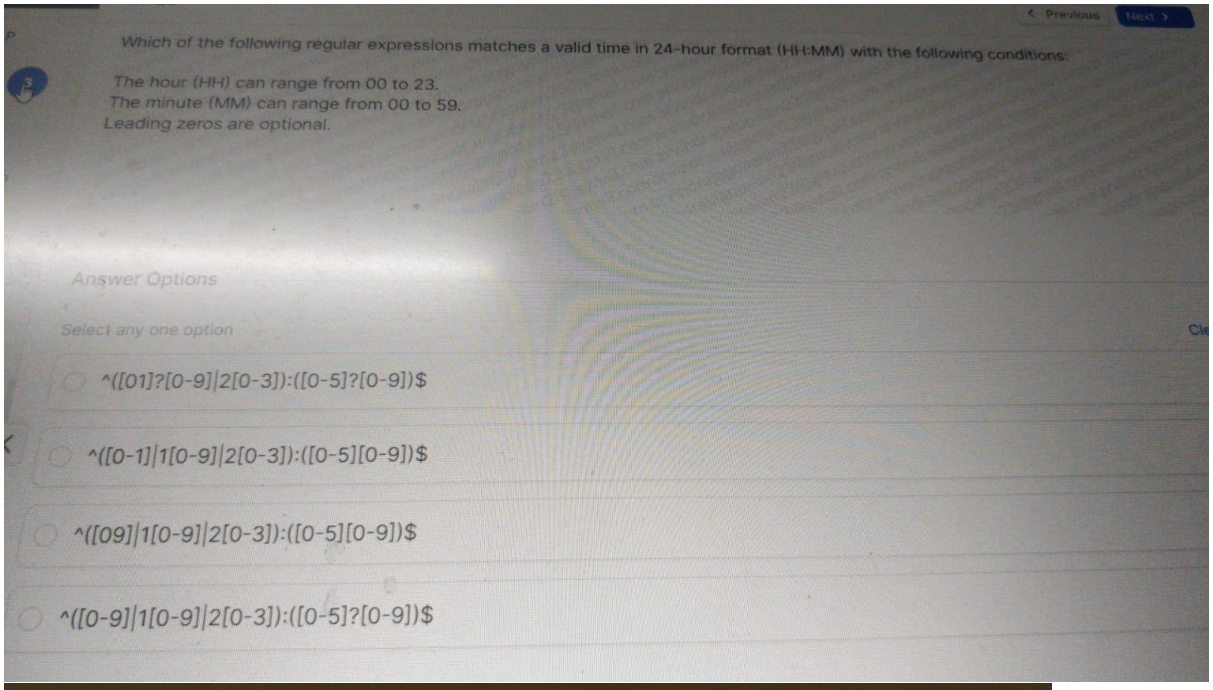
Answer Options

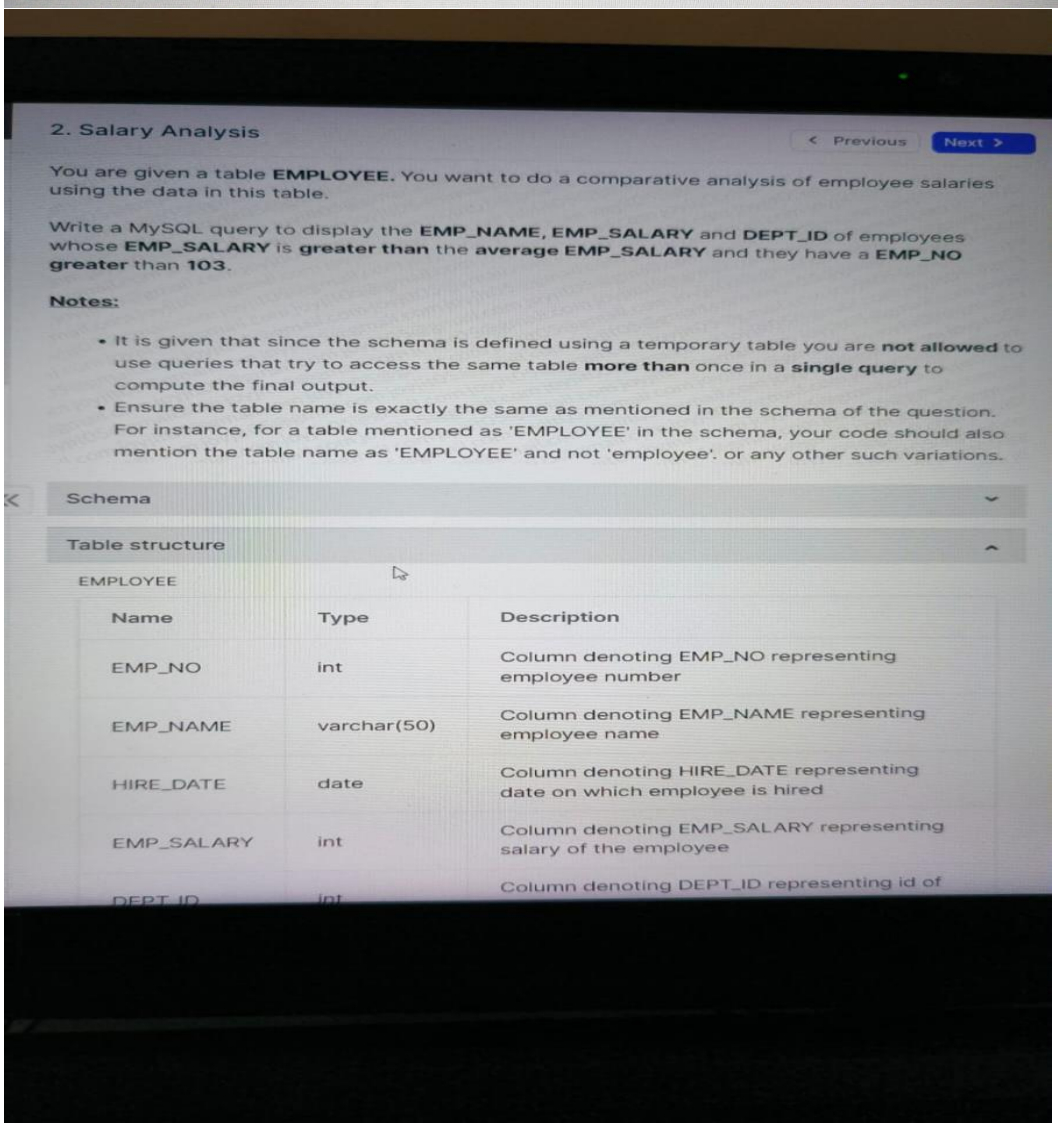
Select any one option

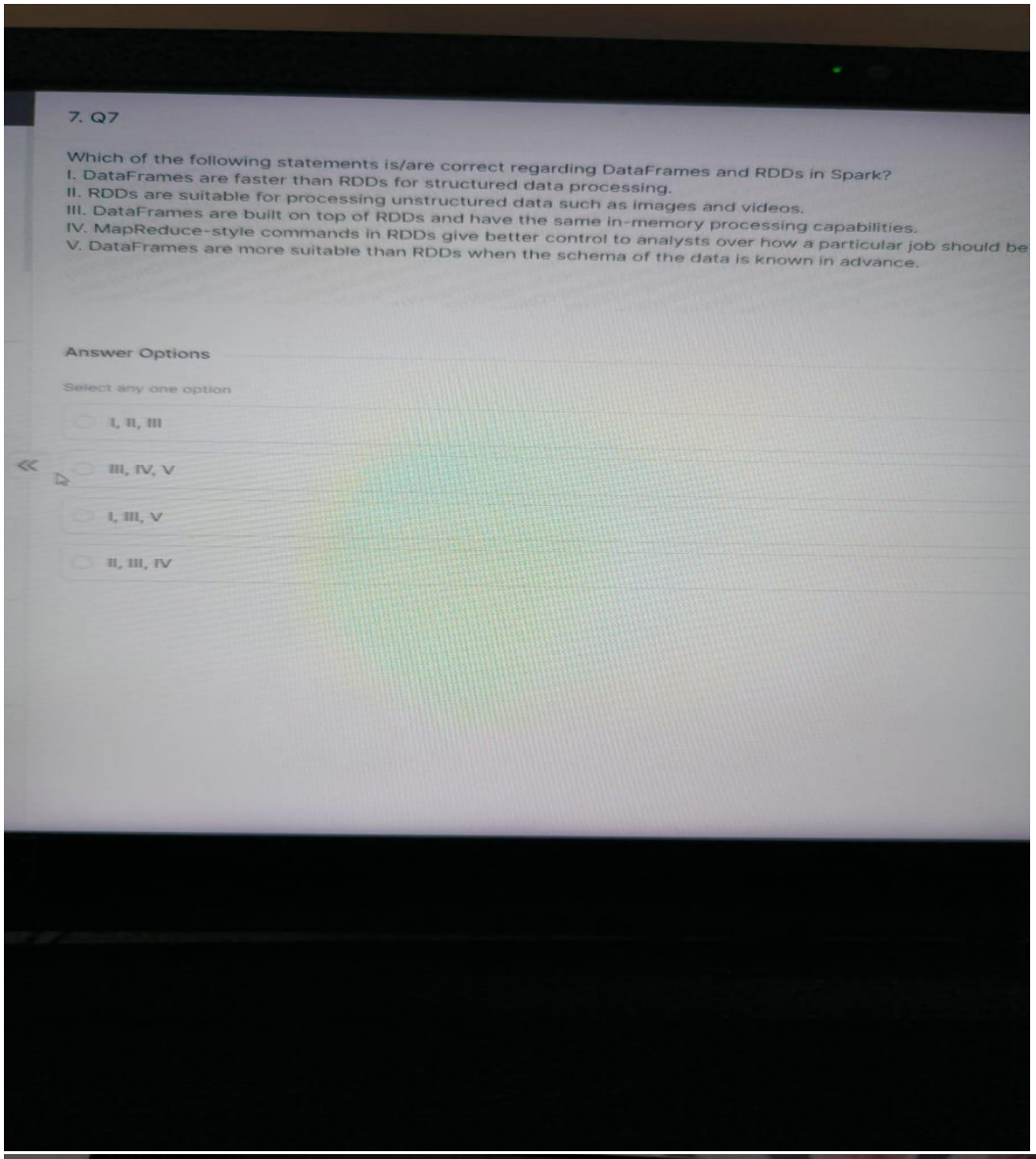
Clear Answer

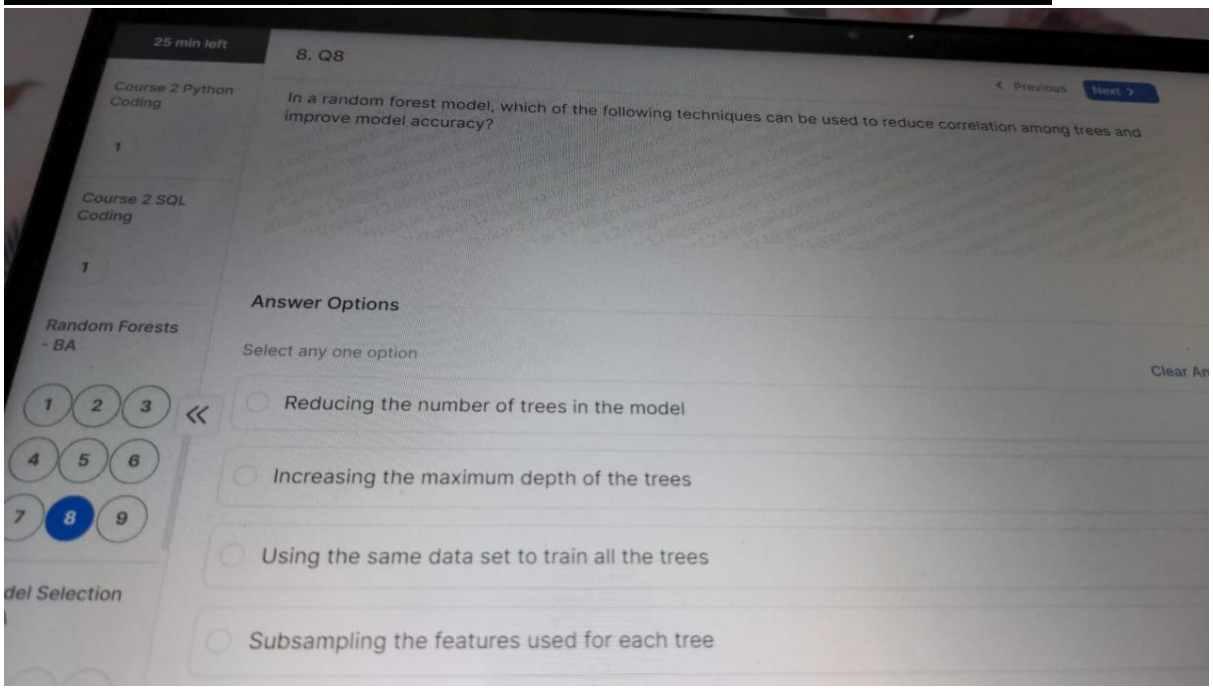
- Neural network
- Decision trees
- Random Forest
- All of the above are equally preferred

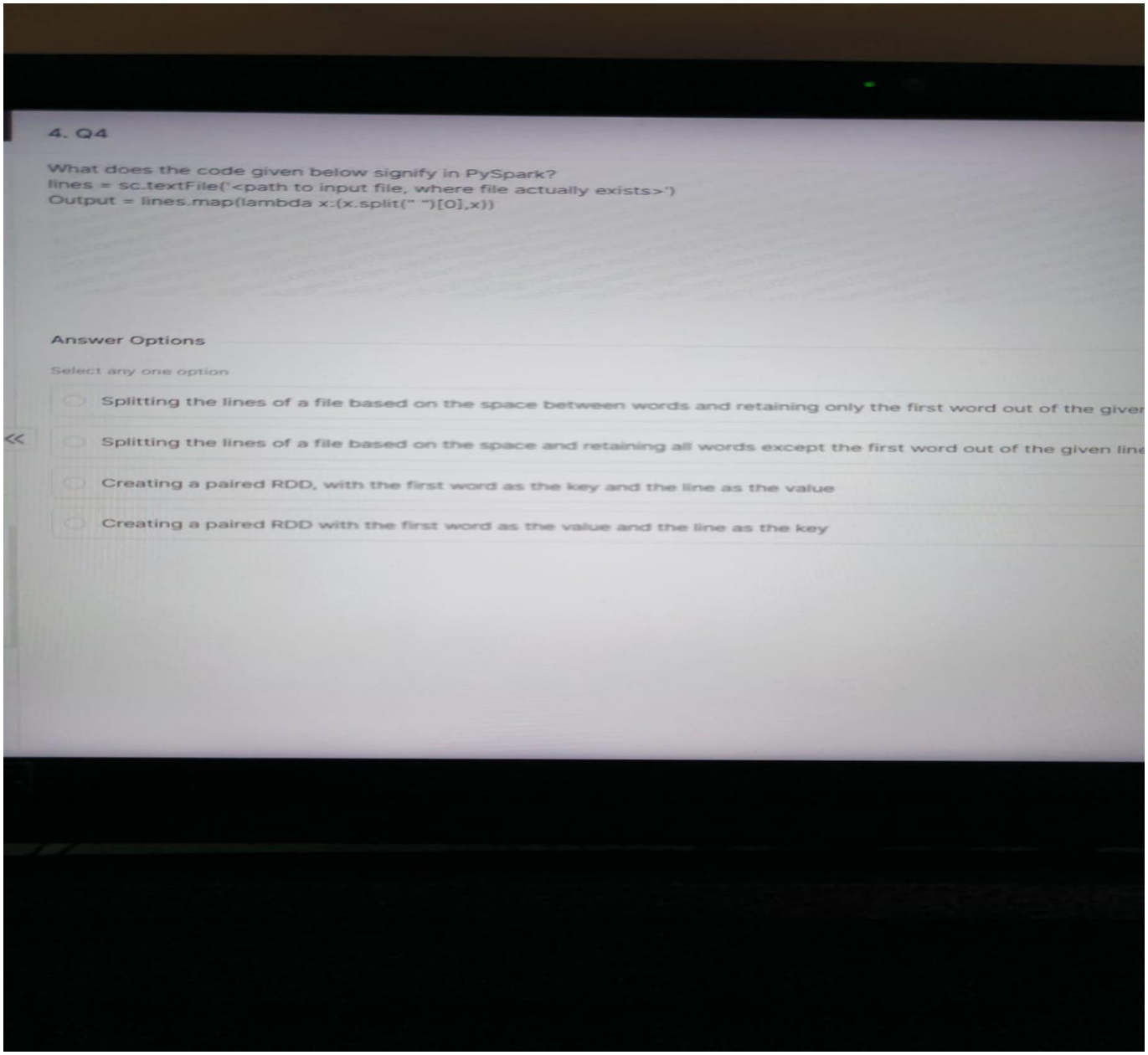



Previous
Next
Which of the following regular expressions matches a valid time in 24-hour format (HH:MM) with the following conditions:
The hour (HH) can range from 00 to 23.
The minute (MM) can range from 00 to 59.
Leading zeros are optional.
Answer Options
Select any one option
^([01]?[0-9]|2[0-3]):([0-5]?[0-9])$
^([0-1]|1[0-9]|2[0-3]):([0-5][0-9])$
^([09]|1[0-9]|2[0-3]):([0-5][0-9])$
^([0-9]|1[0-9]|2[0-3]):([0-5]?[0-9])$


2. 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.
Ensure the table name is exactly the same as mentioned in the schema of the question. For instance, for a table mentioned as 'EMPLOYEE' in the schema, your code should also mention the table name as 'EMPLOYEE' and not 'employee'. or any other such variations.
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


7. Q7
Which of the following statements is/are correct regarding DataFrames and RDDs in Spark?
I. DataFrames are faster than RDDs for structured data processing.
II. RDDs are suitable for processing unstructured data such as images and videos.
III. DataFrames are built on top of RDDs and have the same in-memory processing capabilities.
IV. MapReduce-style commands in RDDs give better control to analysts over how a particular job should be
V. DataFrames are more suitable than RDDs when the schema of the data is known in advance.
Answer Options
Select any one option
I, II, III
III, IV, V
I, III, V
II, III, IV


25 min left
8. Q8
Previous
Next
Course 2 Python Coding
1
Course 2 SQL Coding
1
Random Forests - BA
In a random forest model, which of the following techniques can be used to reduce correlation among trees and improve model accuracy?
Answer Options
Select any one option
Clear An
Reducing the number of trees in the model
Increasing the maximum depth of the trees
Using the same data set to train all the trees
Subsampling the features used for each tree
del Selection



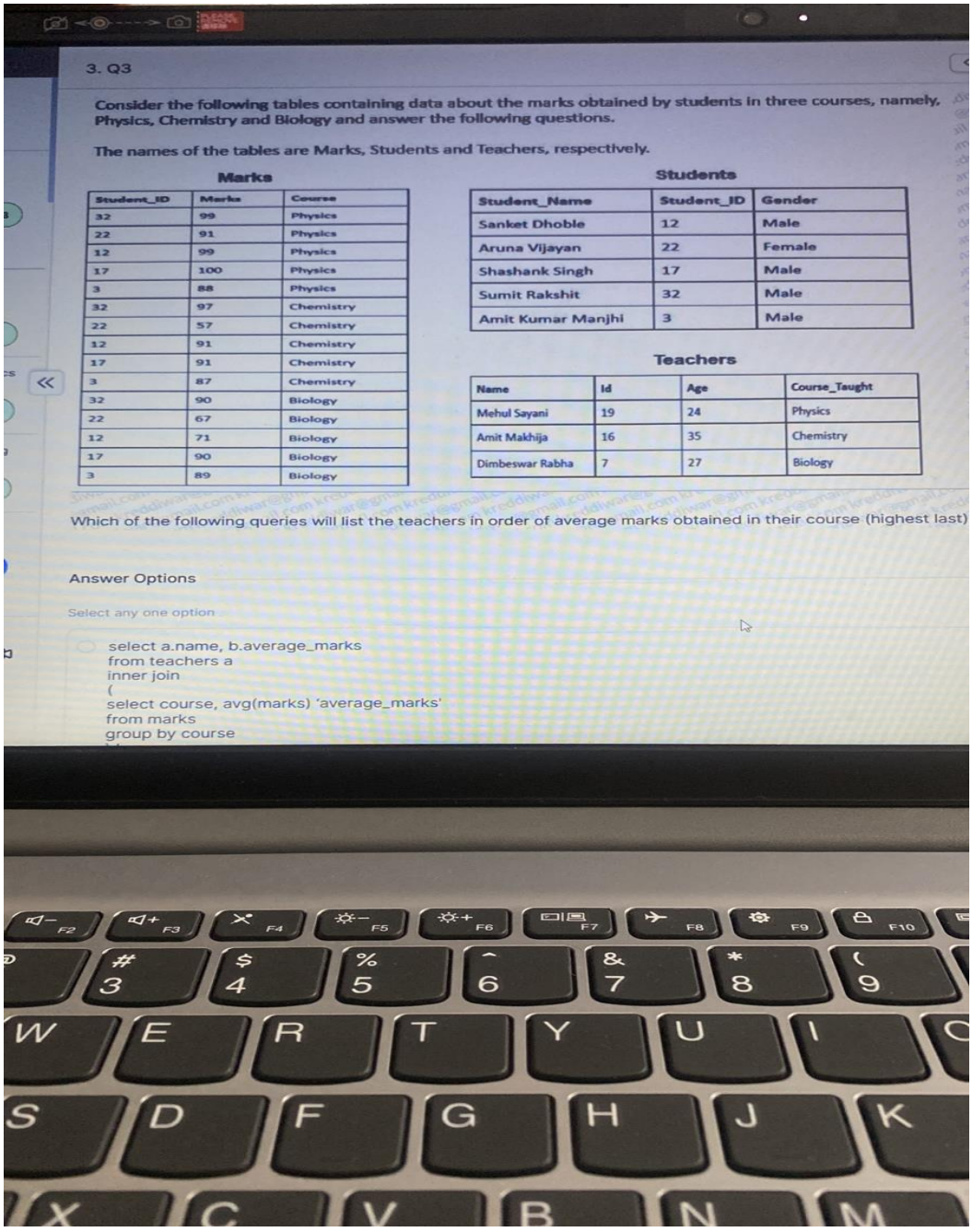
4. Q4

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

- Splitting the lines of a file based on the space between words and retaining only the first word out of the given
- 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



3. Q3

Consider the following tables containing data about the marks obtained by students in three courses, namely, Physics, Chemistry and Biology and answer the following questions.

The names of the tables are Marks, Students and Teachers, respectively.

Marks

Student_ID	Marks	Course
32	99	Physics
22	91	Physics
12	99	Physics
17	100	Physics
3	88	Physics
32	97	Chemistry
22	57	Chemistry
12	91	Chemistry
17	91	Chemistry
3	87	Chemistry
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
Amit Makhija	16	35	Chemistry
Dimbeswar Rabha	7	27	Biology

Which of the following queries will list the teachers in order of average marks obtained in their course (highest last)

Answer Options

Select any one option

```
select a.name, b.average_marks
from teachers a
inner join
(
select course, avg(marks) 'average_marks'
from marks
group by course
```

5. Q5

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

Answer Options

Select any one option

- Map
- Filter
- FlatMap
- ReduceByKey

2. Q2

Which of the following is the correct method to convert an RDD to a DataFrame in PySpark?

Answer Options

Select any one option

- RDD.createDF()
- RDD.convertDF()
- RDD.toDF()
- RDD.asDataFrame()

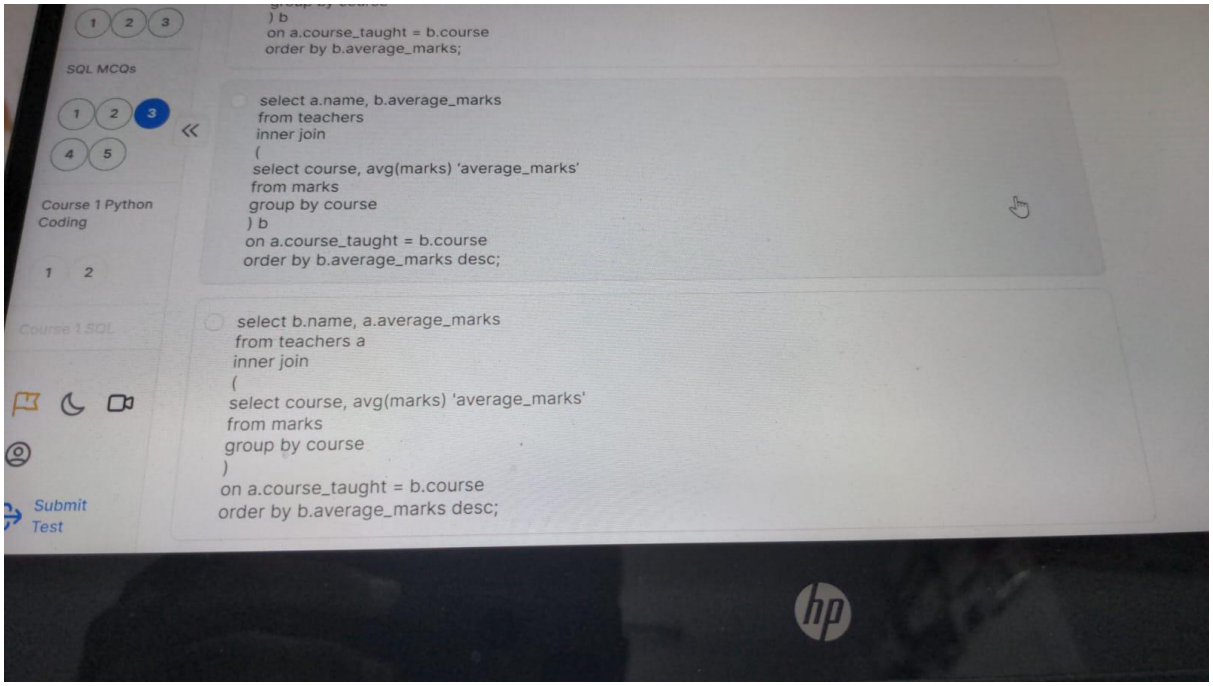
1. Q1

You are working on a Spark program that involves multiple operations on a large DataFrame, and you want to op
Which of the following strategies would you use to achieve this goal?

Answer Options

Select any one option

- Increase the number of partitions of the DataFrame to utilize more resources.
- Use a smaller cluster to reduce the network traffic.
- Use a distributed cache to store the intermediate results of the DataFrame transformations.
- Use a smaller block size for HDFS to improve data locality.
- Use the reduceByKey function to merge the data in the DataFrame and reduce the number of rows.


1 2 3
SQL MCQs
1 2 3
4 5
Course 1 Python Coding
1 2
Course 1 SQL
Submit Test
) b
on a.course_taught = b.course
order by b.average_marks;
select a.name, b.average_marks
from teachers
inner join
(
select course, avg(marks) 'average_marks'
from marks
group by course
) b
on a.course_taught = b.course
order by b.average_marks desc;
select b.name, a.average_marks
from teachers a
inner join
(
select course, avg(marks) 'average_marks'
from marks
group by course
)
on a.course_taught = b.course
order by b.average_marks desc;

1. 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** 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 yo 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.
- Ensure the table name is exactly the same as mentioned in the schem of the question. For instance, for a table mentioned as 'EMPLOYEE' in the schema, your code should also mention the table name as 'EMPLOYEE' and not 'employee'. or any other such variations.

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

3. Q3

Which of the following is true about Spark's caching mechanism?

Answer Options

Select any one option

- Caching is used to store the DataFrame on the local file system.
- Caching improves performance by reducing I/O operations and recomputation.
- Caching requires additional disk space to store the cached data.
- Caching cannot be used in Spark Streaming applications.

Hypothesis Testing

1 2 3

SQL MCQs

1 2 3

4 5

urse 1 Python ing

2

1 SQL

Marks

Student_ID	Marks	Course
32	99	Physics
22	91	Physics
12	99	Physics
17	100	Physics
3	88	Physics
32	97	Chemistry
22	57	Chemistry
12	91	Chemistry
17	91	Chemistry
3	87	Chemistry
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
Amit Makhija	16	35	Chemistry
Dimbeswar Rabha	7	27	Biology

Which of the following queries will list the teachers in order of average marks obtained in their course (highest last)?

Answer Options

1. Salary Analysis

Previous | Next

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

ıut

PLOYEE

EMP_NO	EMP_NAME	HIRE_DATE	EMP_SALARY	DEPT_
103	Vipul	1990-10-11	5000	34
104	John	2020-11-11	3000	15
105	Ram	2020-10-11	10000	34

tput

Ram	10000	34

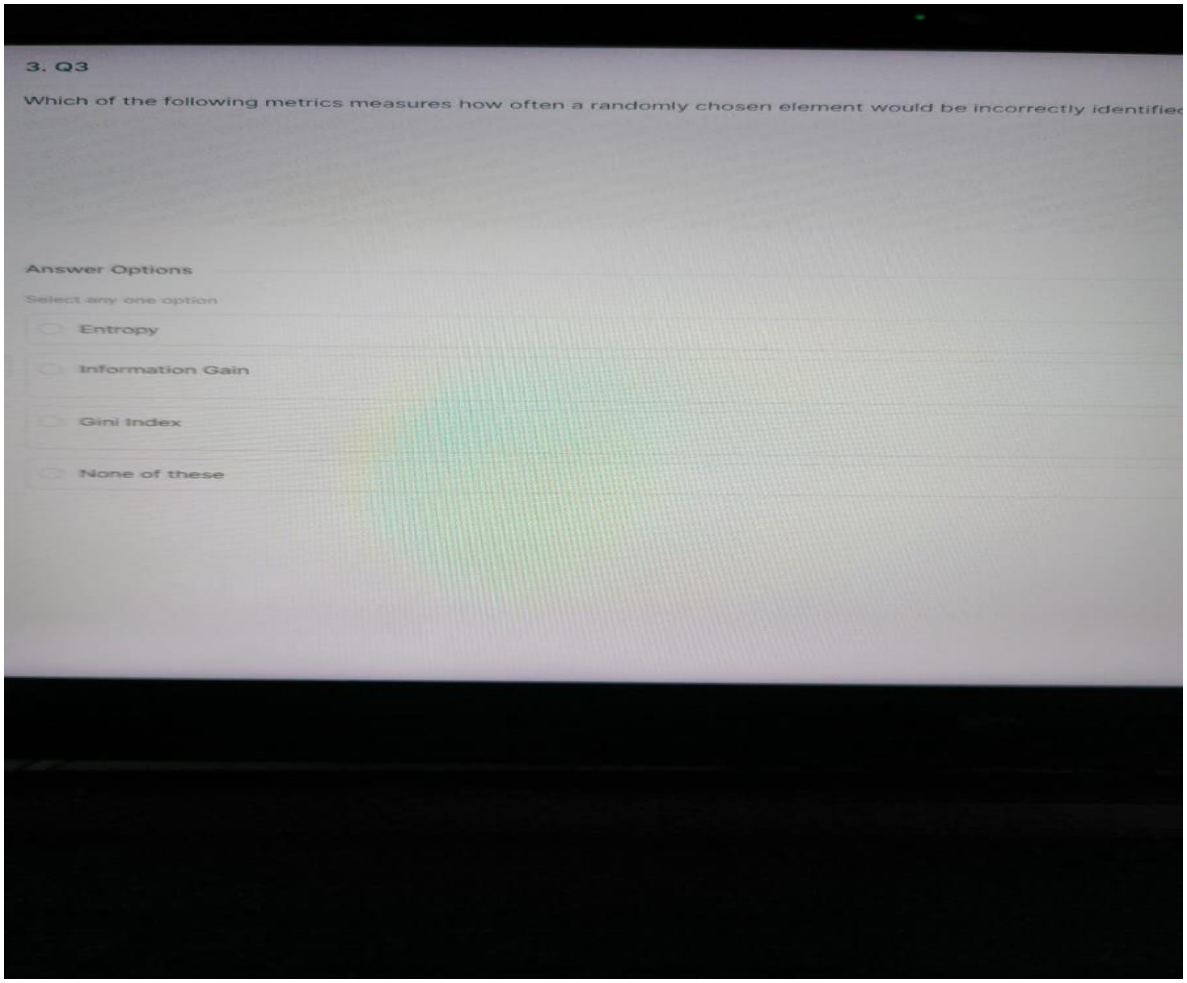
3. Q3

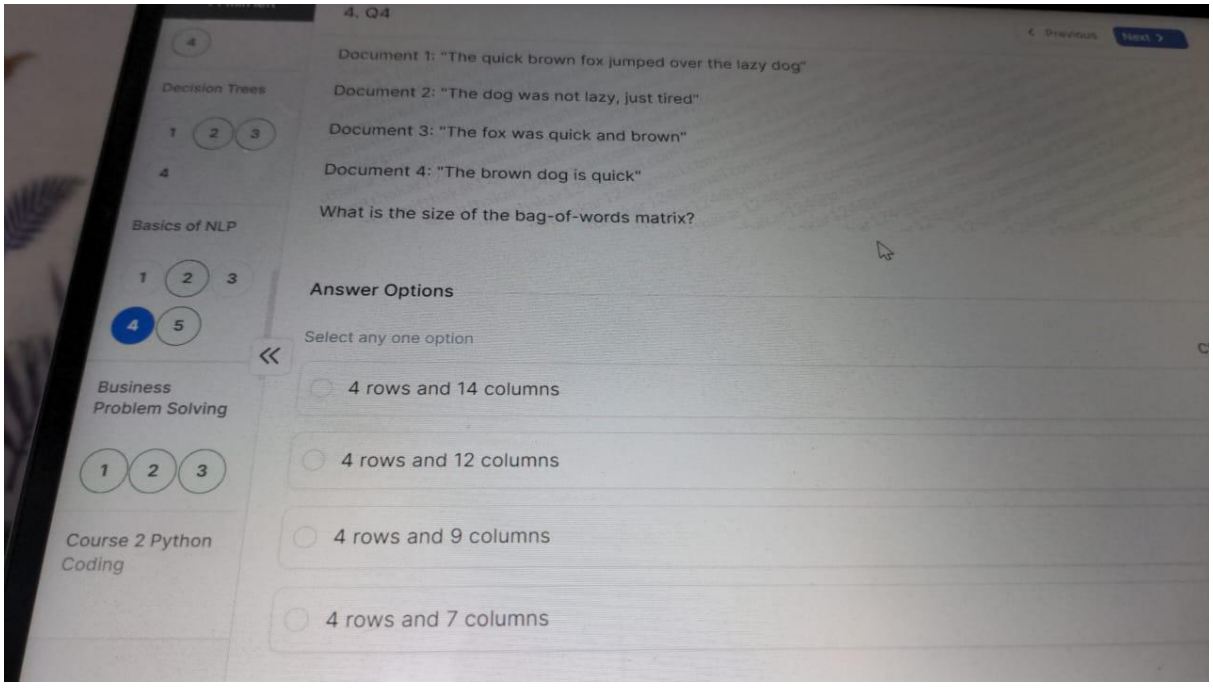
Which of the following metrics measures how often a randomly chosen element would be incorrectly identifie

Answer Options

Select any one option

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


3. Q3
Which of the following metrics measures how often a randomly chosen element would be incorrectly identifie
Answer Options
Select any one option
Entropy
Information Gain
Gini Index
None of these


4
Decision Trees
1 2 3
4
Basics of NLP
1 2 3
4 5
Business Problem Solving
1 2 3
Course 2 Python Coding
4. Q4
Previous
Next
Document 1: "The quick brown fox jumped over the lazy dog"
Document 2: "The dog was not lazy, just tired"
Document 3: "The fox was quick and brown"
Document 4: "The brown dog is quick"
What is the size of the bag-of-words matrix?
Answer Options
Select any one option
4 rows and 14 columns
4 rows and 12 columns
4 rows and 9 columns
4 rows and 7 columns

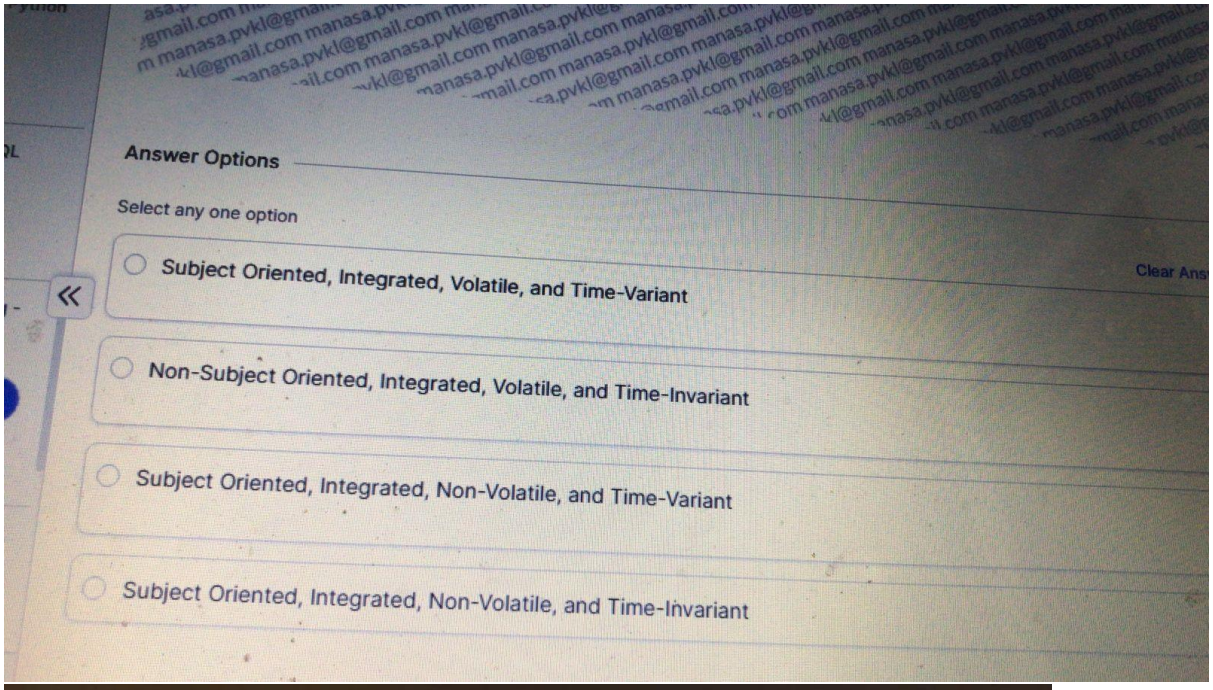
5. Q5

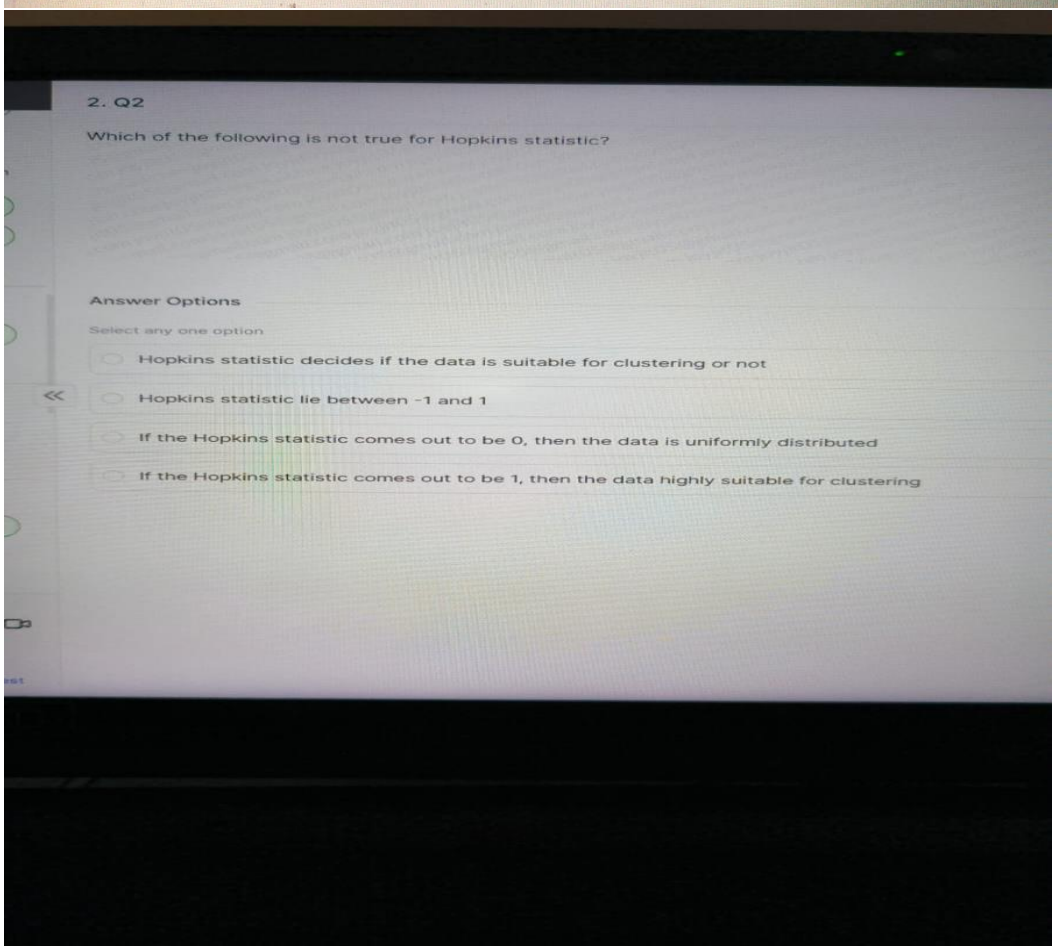
If you use a random number generator to predict the output 0 or 1 for a binary classification problem, what will be the area u
of the ROC curve?

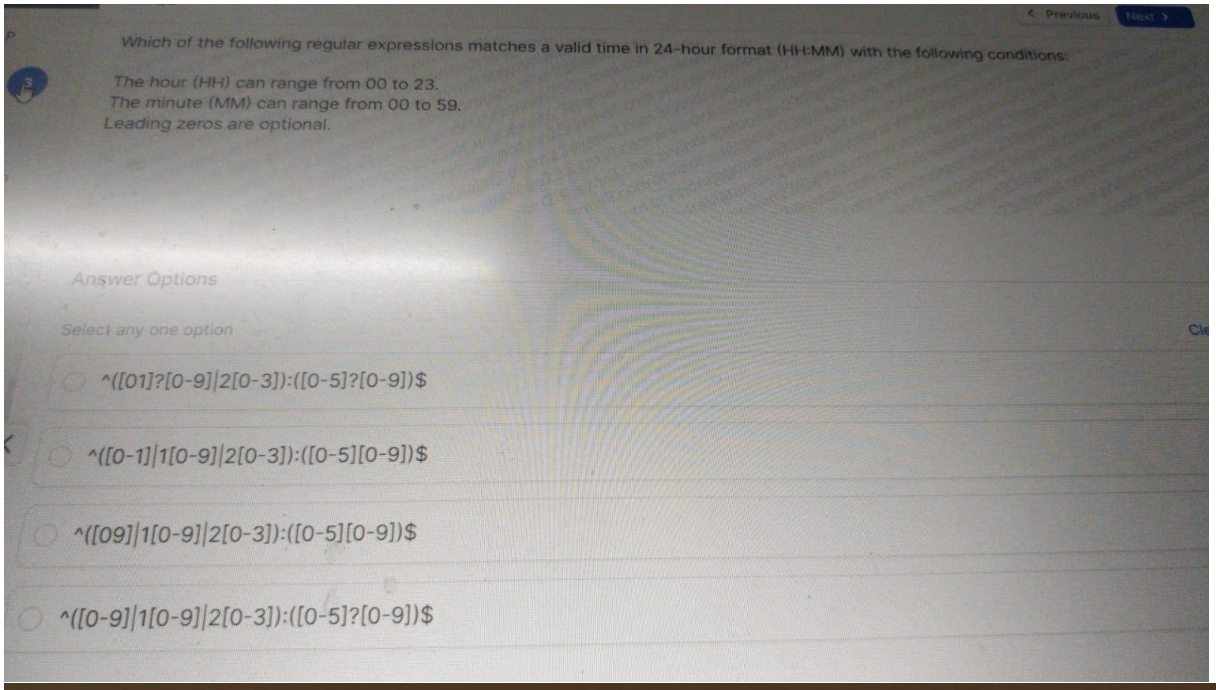
Answer Options

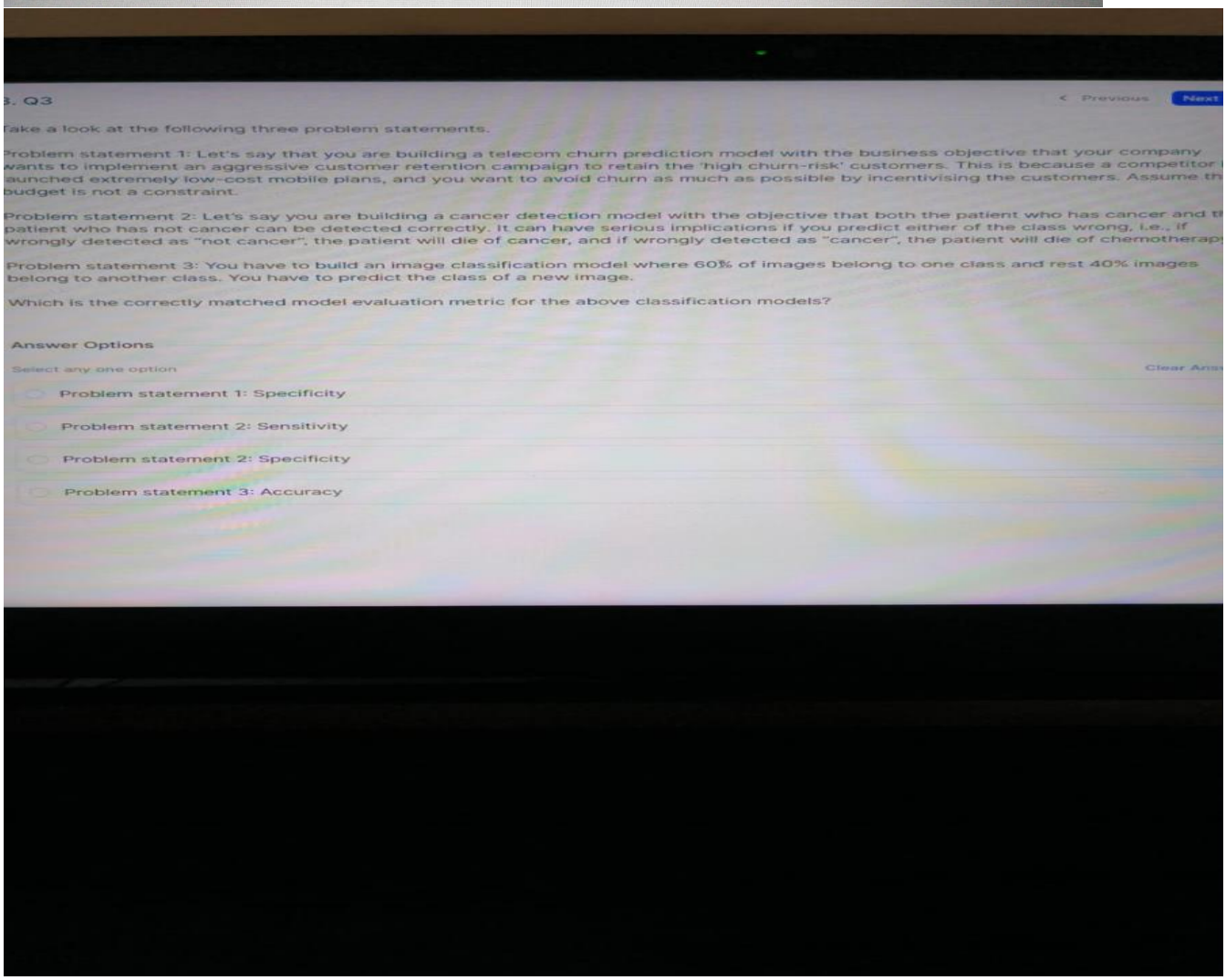
Select any one option

- 0
- 0.5
- 1
- 100


Answer Options
Select any one option
Subject Oriented, Integrated, Volatile, and Time-Variant
Non-Subject Oriented, Integrated, Volatile, and Time-Invariant
Subject Oriented, Integrated, Non-Volatile, and Time-Variant
Subject Oriented, Integrated, Non-Volatile, and Time-Invariant


2. Q2
Which of the following is not true for Hopkins statistic?
Answer Options
Select any one option
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


Previous
Next
Which of the following regular expressions matches a valid time in 24-hour format (HH:MM) with the following conditions:
The hour (HH) can range from 00 to 23.
The minute (MM) can range from 00 to 59.
Leading zeros are optional.
Answer Options
Select any one option
^([01]?[0-9]|2[0-3]):([0-5]?[0-9])$
^([0-1]|1[0-9]|2[0-3]):([0-5][0-9])$
^([09]|1[0-9]|2[0-3]):([0-5][0-9])$
^([0-9]|1[0-9]|2[0-3]):([0-5]?[0-9])$


Q3
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 launched extremely low-cost mobile plans, and you want to avoid churn as much as possible by incentivising the customers. Assume th 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 t 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 chemotherap
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
Problem statement 1: Specificity
Problem statement 2: Sensitivity
Problem statement 2: Specificity
Problem statement 3: Accuracy