Your grade: 100%

Your latest: **100%** • Your highest: **100%** • To pass you need at least 70%. We keep your highest score.

Next item

1. Which is the syntax code to split the data into 60% training data and 40% testing data? 1 / 1 point

- (●) training_data, testing_data = data.randomSplit([0.6, 0.4])
- ( ) training_data, testing_data = data.randomSplit([0.4, 0.6])
- ( ) testing_data, training_data = data.randomSplit([40, 60])
- ( ) testing_data, training_data = data.randomSplit([0.6, 0.4])

Correct
Correct! This is the correct syntax to split the data into 60% training data and 40% testing data.

2. What does a VectorAssembler do? 1 / 1 point

- (●) It combines a bunch of columns as a single vector column.
- ( ) It combines individual data elements into a row.
- ( ) It combines two DataFrames into one.
- ( ) It combines the individual data elements into a column.

Correct
Correct! A VectorAssembler is a transformer that combines the values of the selected feature columns into a single vector.

3. What is the primary purpose of Spark's in-memory processing capability? 1 / 1 point

- (●) To reduce disk-based I/O costs
- ( ) To enable real-time data stream processing
- ( ) To improve data ingestion performance
- ( ) To support complex data transformation tasks

Correct
Correct! Spark's in-memory processing capability helps reduce the high I/O costs associated with traditional disk-based processing systems.

4. What is the role of data engineers in Spark cluster monitoring? 1 / 1 point

- (●) To ensure the efficient running and health of the Spark cluster
- ( ) To troubleshoot issues related to data ingestion pipelines
- ( ) To optimize code and data structures for better performance
- ( ) To analyze and visualize data processed by Spark

Correct
Correct! Data engineers play a crucial role in monitoring the Spark cluster to ensure it is running efficiently, detecting, and resolving any issues that may arise.

5. Your goal is to predict the height of a child, given the age and the weight. Which of the following algorithms will help you achieve that? 1 / 1 point

- (●) Linear regression
- ( ) K-means
- ( ) Logistic regression
- ( ) RandomSplit

Correct
Correct! The linear regression algorithm helps to create a model that can be used to predict a numerical value.

6. Which is the correct statement for a linear regression problem? 1 / 1 point

- ( ) There will be 1 label column, which is non-numeric and multiple numeric feature columns.
- (●) There will be 1 label column, which is numeric and multiple numeric feature columns.
- ( ) There will be 1 label column, which is text and multiple numeric feature columns.
- ( ) There will be 1 label column, which is non-numeric and multiple non-numeric feature columns.

Correct
Correct! There will be 1 label column, which is numeric and multiple numeric feature columns.

7. Which is the correct syntax to create a Spark session with application name "Test App"? 1 / 1 point

- spark = SparkSession.builder.appname("Test App").createSession()
- spark = SparkSession.builder.appName("Test App").getOrCreate()
- spark = SparkSession.builder.appname("Test App").getOrCreate
- spark = Sparksession.builder.appName("Test App").getOrCreateSession()

Correct
Correct! The correct syntax is spark = SparkSession.builder.appName("Test App").getOrCreate().

8. Which statement best defines Clustering using Spark ML? 1 / 1 point

- It discovers patterns and structures based on their randomness.
- It relies on predefined labels or target variables.
- It is the process of grouping similar data points together into clusters.
- It is a supervised learning technique.

Correct
Correct! It is the process of grouping similar data points together into clusters based on their intrinsic characteristics or patterns.

9. Which is the correct syntax to display the columns "height" and "weight" from the dataframe named "health"? 1 / 1 point

- health.show(["height","weight"])
- health.show("height","weight")
- health.selectcolumns("height","weight").show()
- health.select(["height","weight"]).show()

Correct
Correct !

10. Which statement best defines GraphFrames? 1 / 1 point

- GraphFrames is an integral part of the Spark installation and need not be downloaded as a separate package.
- GraphFrames enables Spark to perform graph processing, run computations, and analyze standard graphs.
- GraphFrames does not require setting a directory for checkpoints.
- GraphFrames does not contain any built-in algorithms; you can download them as a separate package as per your requirements.

Correct
Correct! GraphFrames is an extension to Apache Spark that enables Spark to perform graph processing, run computations, and analyze standard graphs.