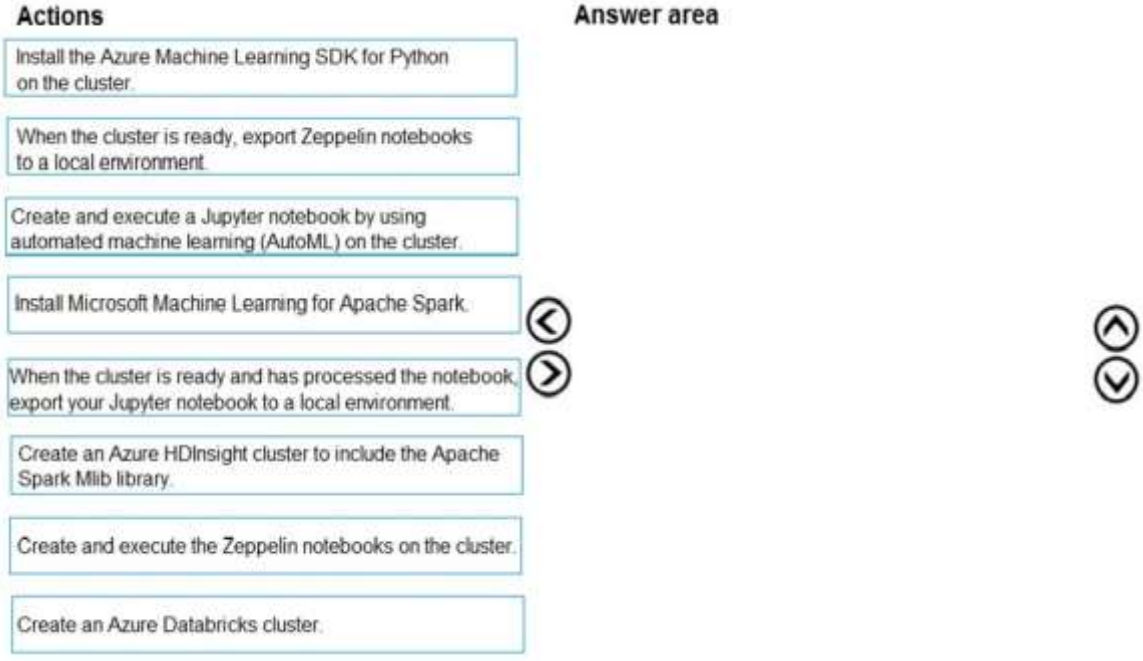
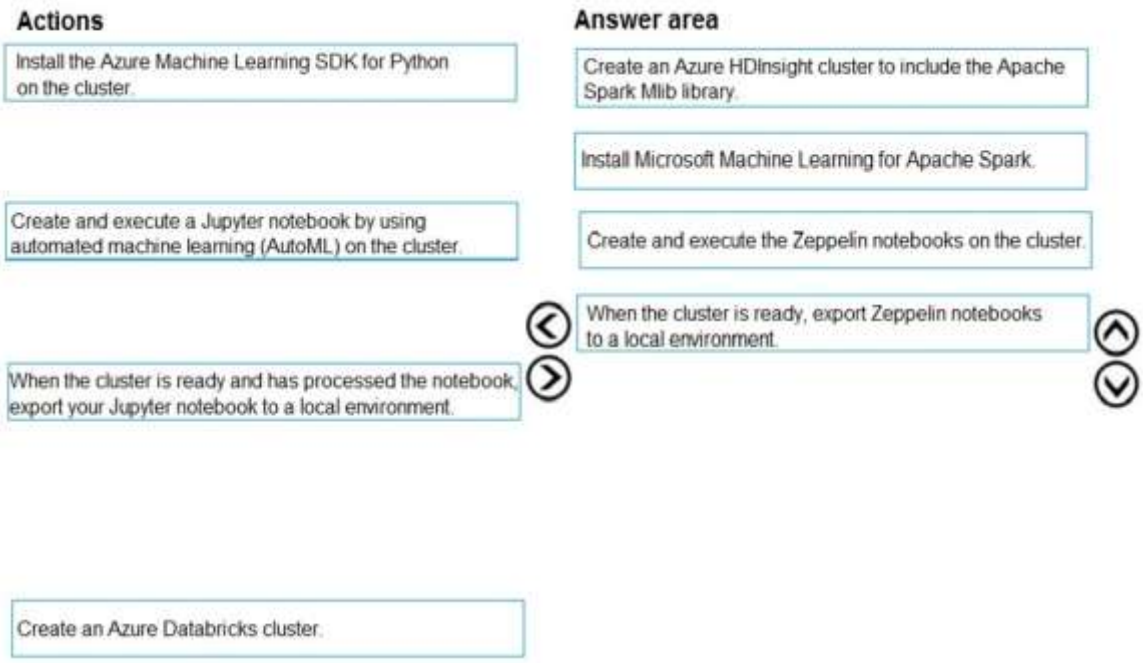
- **Vendor: Microsoft**
- **Exam Code: DP-100**
- **Exam Name: Designing and Implementing a Data Science Solution on Azure**
- **New Updated Questions from Braindump2go (Updated in May/2020)**

QUESTION 61
Drag and Drop Question
You are building an intelligent solution using machine learning models.
The environment must support the following requirements:
- Data scientists must build notebooks in a cloud environment
- Data scientists must use automatic feature engineering and model building in machine learning pipelines.
- Notebooks must be deployed to retrain using Spark instances with dynamic worker allocation.
- Notebooks must be exportable to be version controlled locally.

You need to create the environment.
Which four actions should you perform in sequence? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.



Answer:

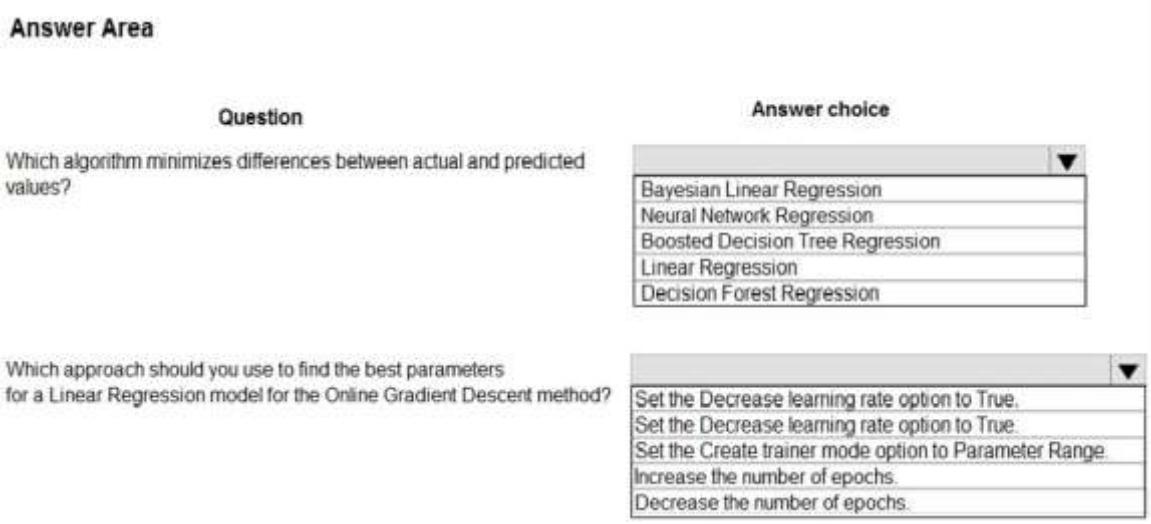


Explanation:
Step 1: Create an Azure HDInsight cluster to include the Apache Spark Mlib library
Step 2: Install Microsot Machine Learning for Apache Spark
You install AzureML on your Azure HDInsight cluster.
Microsoft Machine Learning for Apache Spark (MMLSpark) provides a number of deep learning and data science tools for Apache Spark, including seamless integration of Spark Machine Learning pipelines with Microsoft Cognitive Toolkit (CNTK) and OpenCV, enabling you to quickly create powerful, highly-scalable predictive and analytical models for large image and text datasets.
Step 3: Create and execute the Zeppelin notebooks on the cluster
Step 4: When the cluster is ready, export Zeppelin notebooks to a local environment.
Notebooks must be exportable to be version controlled locally.
References:
https://docs.microsoft.com/en-us/azure/hdinsight/spark/apache-spark-zeppelin-notebook
https://azuremlbuild.blob.core.windows.net/pysparkapi/intro.html

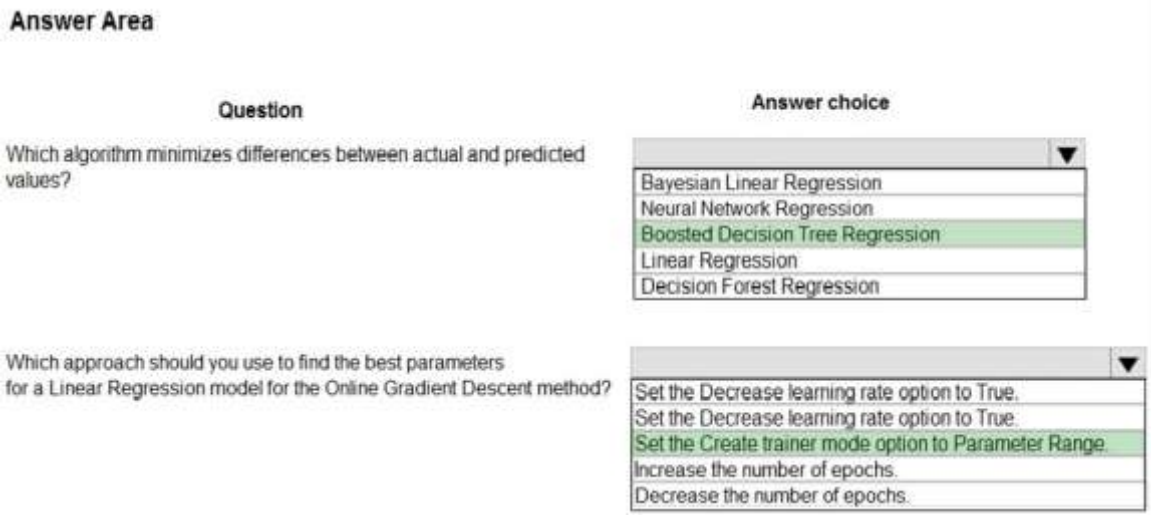
QUESTION 62
Hotspot Question
You are developing a linear regression model in Azure Machine Learning Studio. You run an experiment to compare different algorithms.
The following image displays the results dataset output:

Algorithm	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error	Relative Squared Error
Bayesian Liner	3.276025	4.655442	0.511436	0.282138
Neural Network	2.676538	3.621476	0.417847	0.17073
Boosted Decision Tree	2.168847	2.878077	0.338589	0.107831
Linear	6.350005	8.720718	0.99133	0.99002
Decision Forest	2.390206	3.315 164	0.373146	0.14307

Use the drop-down menus to select the answer choice that answers each question based on the information presented in the image.
NOTE: Each correct selection is worth one point.



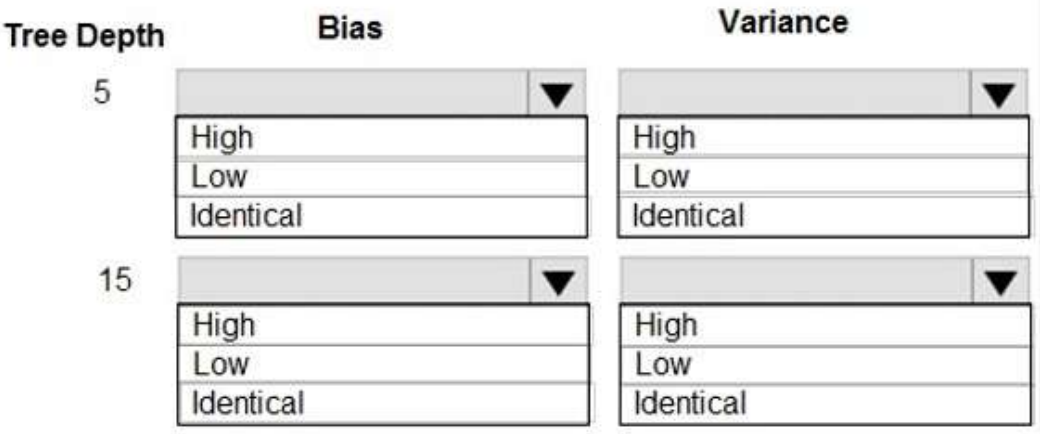
Answer:



Explanation:
Box 1: Boosted Decision Tree Regression
Mean absolute error (MAE) measures how close the predictions are to the actual outcomes; thus, a lower score is better.
Box 2:
Online Gradient Descent: If you want the algorithm to find the best parameters for you, set Create trainer mode option to Parameter Range. You can then specify multiple values for the algorithm to try.
References:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/evaluate-model
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/linear-regression

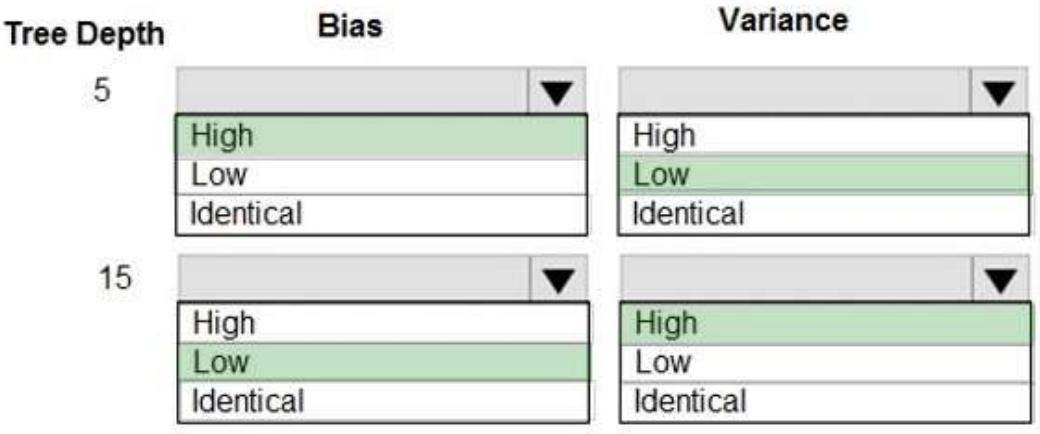
QUESTION 63
Hotspot Question
You are using a decision tree algorithm. You have trained a model that generalizes well at a tree depth equal to 10.
You need to select the bias and variance properties of the model with varying tree depth values.
Which properties should you select for each tree depth? To answer, select the appropriate options in the answer area.

Answer Area



Answer:

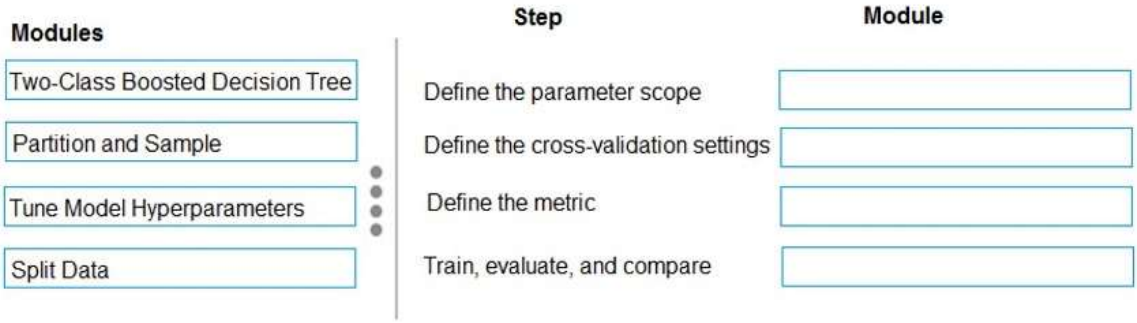
Answer Area



Explanation:
In decision trees, the depth of the tree determines the variance. A complicated decision tree (e.g. deep) has low bias and high variance.
Note: In statistics and machine learning, the bias璿ariance tradeoff is the property of a set of predictive models whereby models with a lower bias in parameter estimation have a higher variance of the parameter estimates across samples, and vice versa. Increasing the bias will decrease the variance. Increasing the variance will decrease the bias.
References:
https://machinelearningmastery.com/gentle-introduction-to-the-bias-variance-trade-off-in-machine-learning/

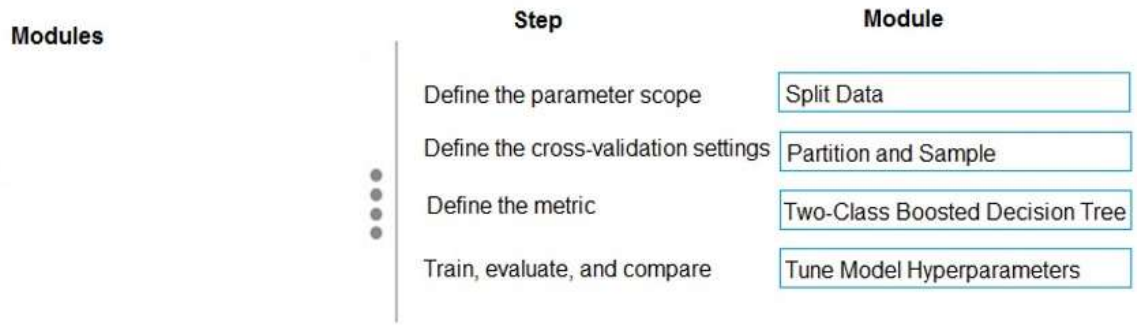
QUESTION 64
Drag and Drop Question
You have a model with a large difference between the training and validation error values.
You must create a new model and perform cross-validation.
You need to identify a parameter set for the new model using Azure Machine Learning Studio.
Which module you should use for each step? To answer, drag the appropriate modules to the correct steps. Each module may be used once or more than once, or not at all. You may need to drag the split bar between panes or scroll to view content.
NOTE: Each correct selection is worth one point.

Answer Area



Answer:

Answer Area



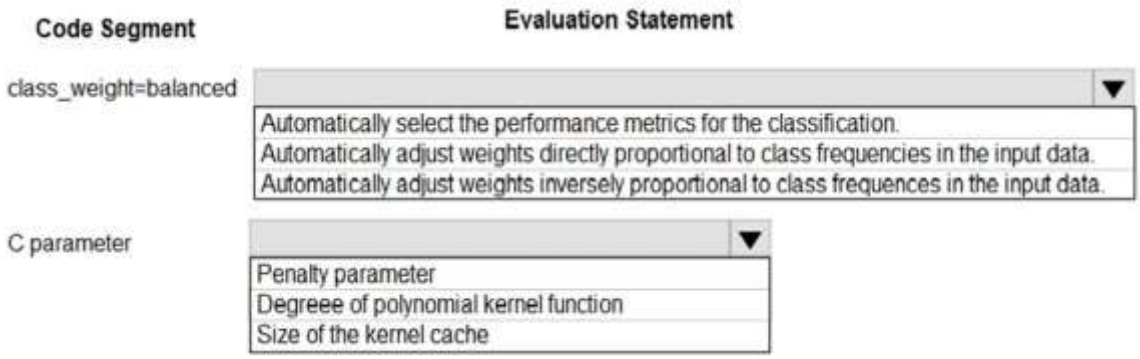
Explanation:
Box 1: Split data
Box 2: Partition and Sample
Box 3: Two-Class Boosted Decision Tree
Box 4: Tune Model Hyperparameters
Integrated train and tune: You configure a set of parameters to use, and then let the module iterate over multiple combinations, measuring accuracy until it finds a "best" model. With most learner modules, you can choose which parameters should be changed during the training process, and which should remain fixed.
We recommend that you use Cross-Validate Model to establish the goodness of the model given the specified parameters. Use Tune Model Hyperparameters to identify the optimal parameters.
References:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/partition-and-sample

QUESTION 65
Hotspot Question
You are using C-Support Vector classification to do a multi-class classification with an unbalanced training dataset. The C-Support Vector classification using Python code shown below:

```
from sklearn.svm import svc
import numpy as np
svc = SVC(kernel= 'linear', class_weight= 'balanced', C-1.0, random_state-0)
model1 = svc.fit(X_train, y)
```

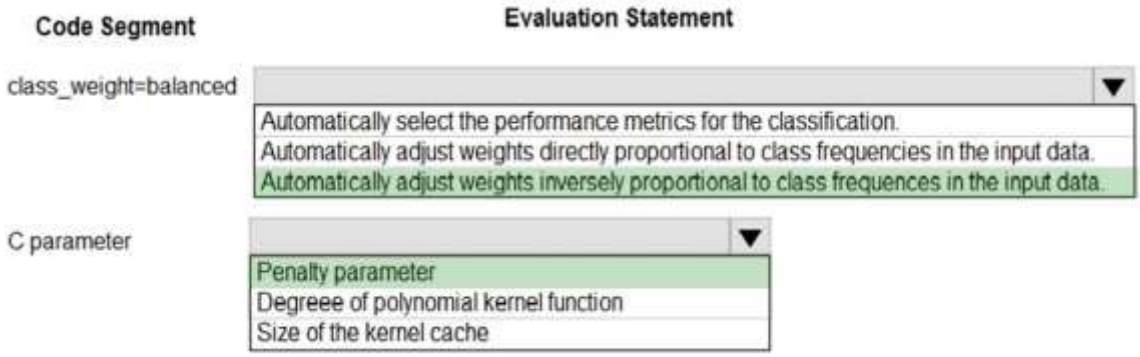
You need to evaluate the C-Support Vector classification code.
Which evaluation statement should you use? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.

Answer Area



Answer:

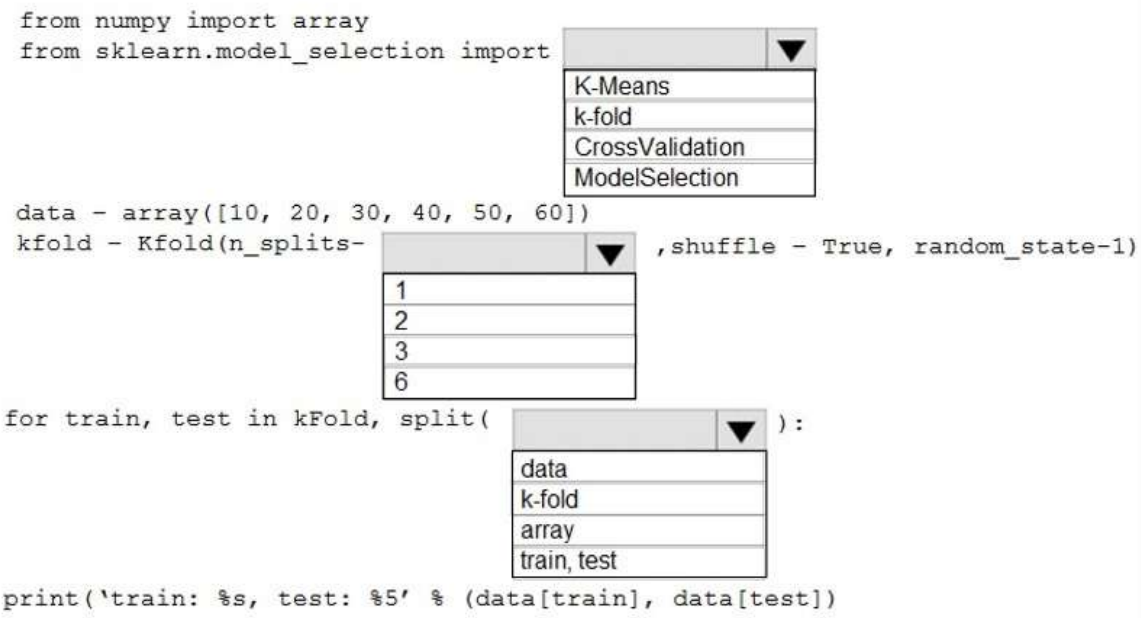
Answer Area



Explanation:
Box 1: Automatically adjust weights inversely proportional to class frequencies in the input data The "balanced" mode uses the values of y to automatically adjust weights inversely proportional to class frequencies in the input data as n_samples / (n_classes * np.bincount(y)).
Box 2: Penalty parameter
Parameter: C : float, optional (default=1.0)
Penalty parameter C of the error term.
References:
https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

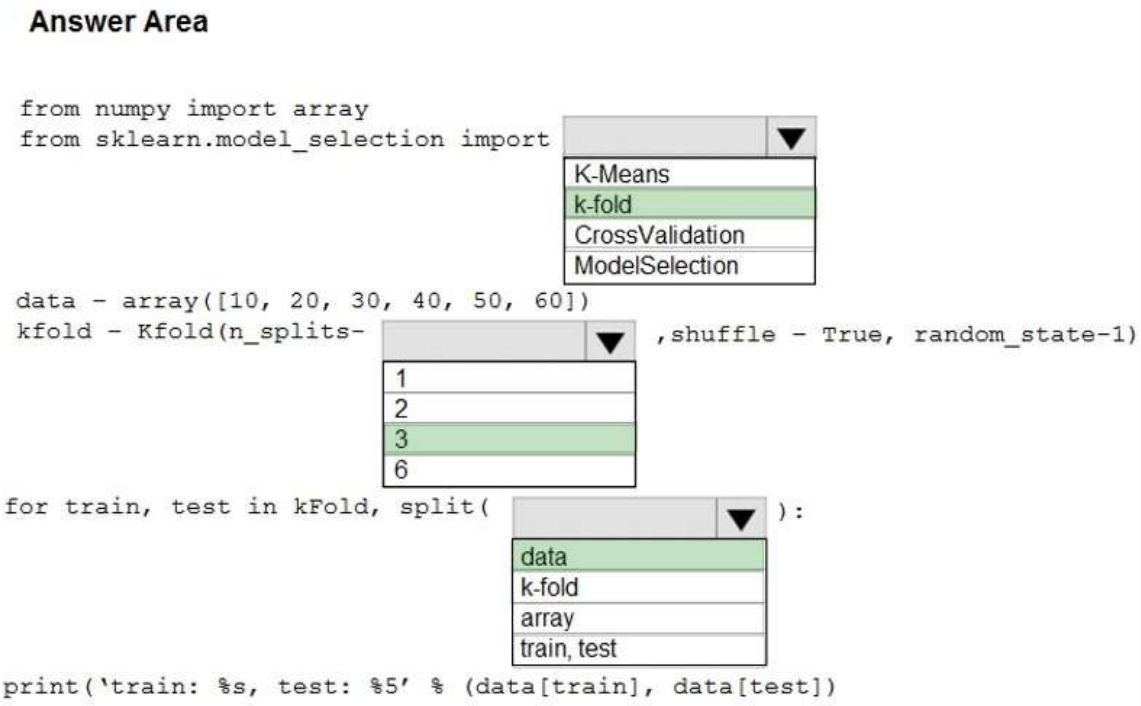
QUESTION 66
Hotspot Question
You are evaluating a Python NumPy array that contains six data points defined as follows:
data = [10, 20, 30, 40, 50, 60]
You must generate the following output by using the k-fold algorithm implantation in the Python Scikit-learn machine learning library:
train: [10 40 50 60], test: [20 30]
train: [20 30 40 60], test: [10 50]
train: [10 20 30 50], test: [40 60]
You need to implement a cross-validation to generate the output.
How should you complete the code segment? To answer, select the appropriate code segment in the dialog box in the answer area.
NOTE: Each correct selection is worth one point.

Answer Area



Answer:

Answer Area



Explanation:

Box 1: k-fold

Box 2: 3

K-Folds cross-validator provides train/test indices to split data in train/test sets. Split dataset into k consecutive folds (without shuffling by default). The parameter n_splits (int, default=3) is the number of folds. Must be at least 2.

Box 3: data

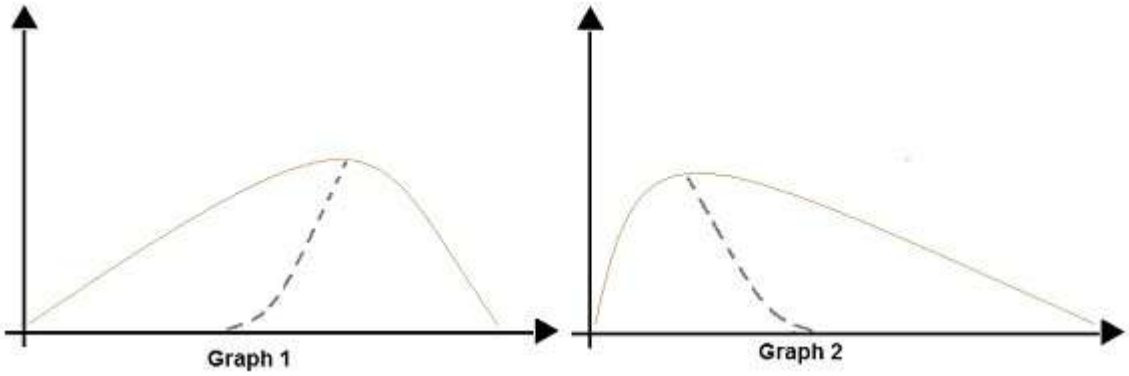
Example: Example:

```
>>>
>>> from sklearn.model_selection import KFold
>>> X = np.array([[1, 2], [3, 4], [1, 2], [3, 4]])
>>> y = np.array([1, 2, 3, 4])
>>> kf = KFold(n_splits=2)
```

```
>>> kf.get_n_splits(X)
>>> print(kf)
KFold(n_splits=2, random_state=None, shuffle=False)
>>> for train_index, test_index in kf.split(X):
... print("TRAIN:", train_index, "TEST:", test_index)
... X_train, X_test = X[train_index], X[test_index]
... y_train, y_test = y[train_index], y[test_index]
TRAIN: [2 3] TEST: [0 1]
TRAIN: [0 1] TEST: [2 3]
```

References:
https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.KFold.html

QUESTION 67
Hotspot Question
You are analyzing the asymmetry in a statistical distribution.
The following image contains two density curves that show the probability distribution of two datasets.



Use the drop-down menus to select the answer choice that answers each question based on the information presented in the graphic.
NOTE: Each correct selection is worth one point.

Answer Area

Question	Answer choice
Which type of distribution is shown for the dataset density curve of Graph 1?	Negative skew / Positive skew / Normal distribution / Bimodal distribution
Which type of distribution is shown for the dataset density curve of Graph 2?	Negative skew / Positive skew / Normal distribution / Bimodal distribution

Answer:

Answer Area

Question	Answer choice
Which type of distribution is shown for the dataset density curve of Graph 1?	Negative skew **Positive skew** Normal distribution Bimodal distribution
Which type of distribution is shown for the dataset density curve of Graph 2?	**Negative skew** Positive skew Normal distribution Bimodal distribution

Explanation:
Box 1: Positive skew
Positive skew values means the distribution is skewed to the right.
Box 2: Negative skew
Negative skewness values mean the distribution is skewed to the left.
References:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/compute-elementary-statistics

QUESTION 68
Hotspot Question
You have a dataset created for multiclass classification tasks that contains a normalized numerical feature set with 10,000 data points and 150 features.
You use 75 percent of the data points for training and 25 percent for testing. You are using the scikit-learn machine learning library in Python. You use X to denote the feature set and Y to denote class labels.
You create the following Python data frames:

Name	Description
X_train	training feature set
Y_train	training class labels
x_train	testing feature set
y_train	testing class labels

You need to apply the Principal Component Analysis (PCA) method to reduce the dimensionality of the feature set to 10 features in both training and testing sets.
How should you complete the code segment? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.

Answer Area

```
from sklearn.decomposition import PCA
pca = [ PCA() | PCA(n_components = 150) | PCA(n_components = 10) | PCA(n_components = 10000) ]
X_train= [ pca | model | sklearn.decomposition ].fit_transform(X_train)
x_test = pca.[ x_test | X_train | fit(x_test) | transform(x_test) ]
```

Answer:

Answer Area

```
from sklearn.decomposition import PCA
pca = [ PCA() | PCA(n_components = 150) | PCA(n_components = 10) | PCA(n_components = 10000) ]
X_train= [ pca | model | sklearn.decomposition ].fit_transform(X_train)
x_test = pca.[ x_test | X_train | fit(x_test) | transform(x_test) ]
```

Explanation:
Box 1: PCA(n_components = 10)

Need to reduce the dimensionality of the feature set to 10 features in both training and testing sets.
Example:
from sklearn.decomposition import PCA
pca = PCA(n_components=2) ;2 dimensions
principalComponents = pca.fit_transform(x)
Box 2: pca
fit_transform(X[, y])fits the model with X and apply the dimensionality reduction on X.
Box 3: transform(x_test)
transform(X) applies dimensionality reduction to X.
References:
https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html

QUESTION 69
Hotspot Question
You have a feature set containing the following numerical features: X, Y, and Z.
The Poisson correlation coefficient (r-value) of X, Y, and Z features is shown in the following image:

	X	Y	Z
X	1	0.149676	-0.106276
Y	0.149676	1	0.859122
Z	-0.106276	0.859122	1

Use the drop-down menus to select the answer choice that answers each question based on the information presented in the graphic.
NOTE: Each correct selection is worth one point.

Answer Area

What is the r-value for the correlation of Y to Z?
-0.106276
0.149676
0.859122
1

Which type of relationship exists between Z and Y in the feature set?
a positive linear relationship
a negative linear relationship
no linear relationship

Answer:

Answer Area

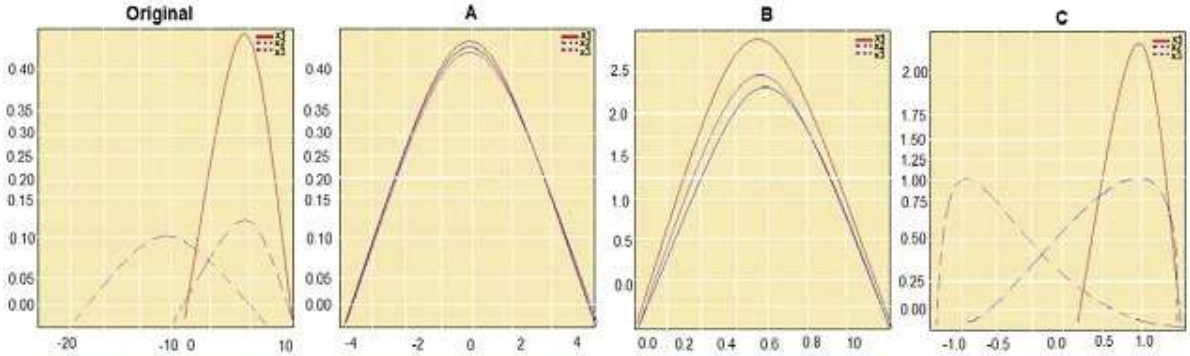
What is the r-value for the correlation of Y to Z?
-0.106276
0.149676
0.859122
1

Which type of relationship exists between Z and Y in the feature set?
a positive linear relationship
a negative linear relationship
no linear relationship

Explanation:
Box 1: 0.859122
Box 2: a positively linear relationship

+1 indicates a strong positive linear relationship
-1 indicates a strong negative linear correlation
0 denotes no linear relationship between the two variables.
References:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/compute-linear-correlation

QUESTION 70
Hotspot Question
You are performing feature scaling by using the scikit-learn Python library for x.1 x2, and x3 features.
Original and scaled data is shown in the following image.



Use the drop-down menus to select the answer choice that answers each question based on the information presented in the graphic.
NOTE: Each correct selection is worth one point.

Answer Area

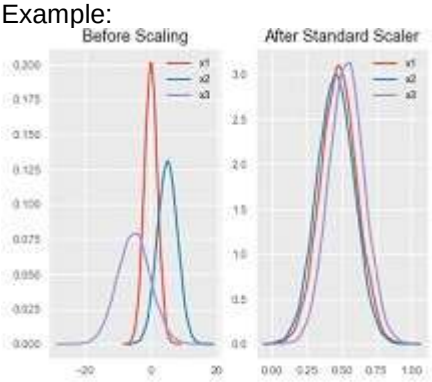
Question	Answer choice
Which scaler is used in graph A?	Standard Scaler / Min Max Scale / Normalizer
Which scaler is used in graph B?	Standard Scaler / Min Max Scale / Normalizer
Which scaler is used in graph C?	Standard Scaler / Min Max Scale / Normalizer

Answer:

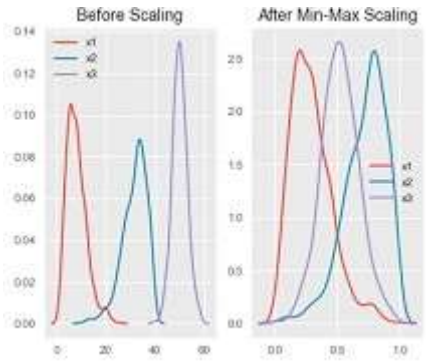
Answer Area

Question	Answer choice
Which scaler is used in graph A?	Standard Scaler Min Max Scale Normalizer
Which scaler is used in graph B?	Standard Scaler Min Max Scale Normalizer
Which scaler is used in graph C?	Standard Scaler Min Max Scale Normalizer

Explanation:
Box 1: StandardScaler
The StandardScaler assumes your data is normally distributed within each feature and will scale them such that the distribution is now centred around 0, with a standard deviation of 1.
Example:



All features are now on the same scale relative to one another.
Box 2: Min Max Scaler



Notice that the skewness of the distribution is maintained but the 3 distributions are brought into the same scale so that they overlap.
Box 3: Normalizer
References:
http://benalexkeen.com/feature-scaling-with-scikit-learn/

QUESTION 71
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.
You are analyzing a numerical dataset which contain missing values in several columns. You must clean the missing values using an appropriate operation without affecting the dimensionality of the feature set.
You need to analyze a full dataset to include all values.
Solution: Use the last Observation Carried Forward (IOCF) method to impute the missing data points.
Does the solution meet the goal?

A. Yes
B. No

Answer: B
Explanation:
Instead use the Multiple Imputation by Chained Equations (MICE) method.
Replace using MICE: For each missing value, this option assigns a new value, which is calculated by using a method described in the statistical literature as "Multivariate Imputation using Chained Equations" or "Multiple Imputation by Chained Equations". With a multiple imputation method, each variable with missing data is modeled conditionally using the other variables in the data before filling in the missing values.
Note: Last observation carried forward (LOCF) is a method of imputing missing data in longitudinal studies. If a person drops out of a study before it ends, then his or her last observed score on the dependent variable is used for all subsequent (i.e., missing) observation points. LOCF is used to maintain the sample size and to reduce the bias caused by the attrition of participants in a study.
References:
https://methods.sagepub.com/reference/encyc-of-research-design/n211.xml
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3074241/