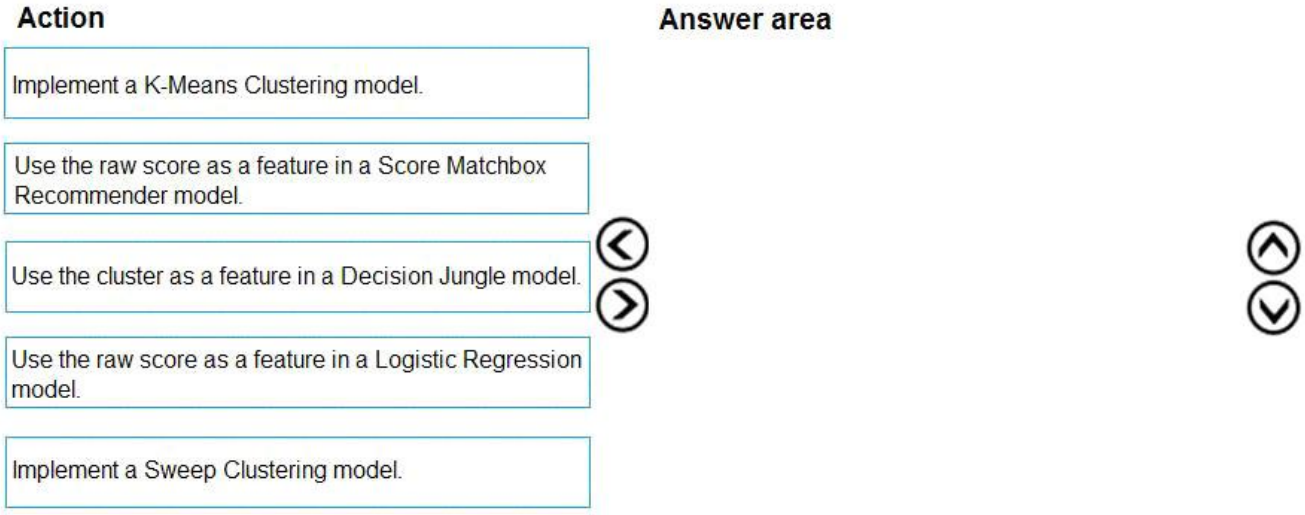
Exam : **DP-100**

Title : Designing and Implementing a Data Science Solution on Azure (beta)

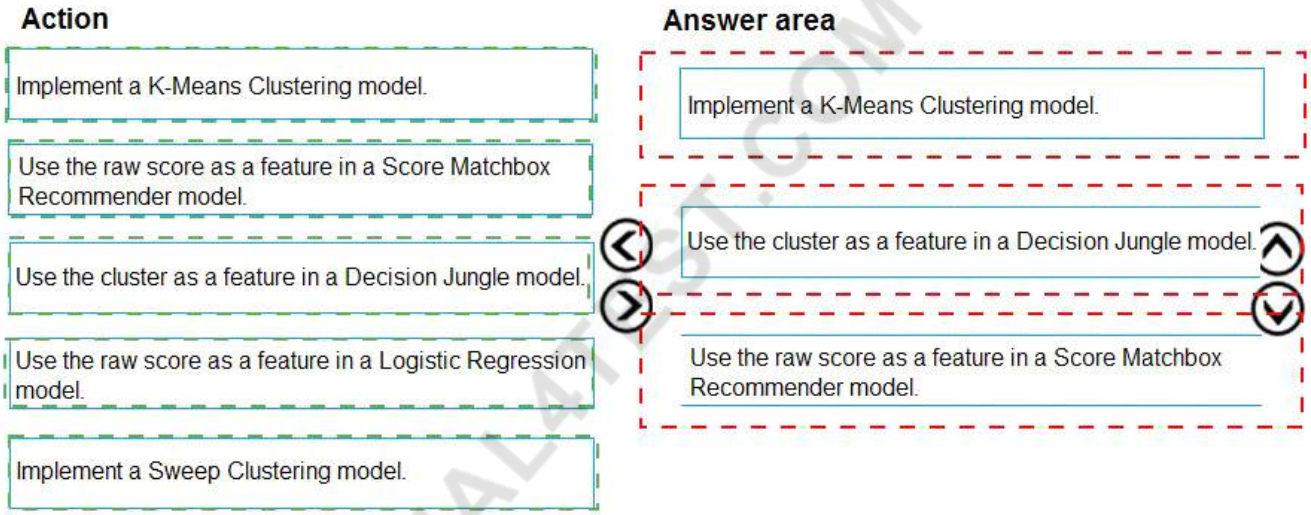
Vendor : Microsoft

Version : V12.75

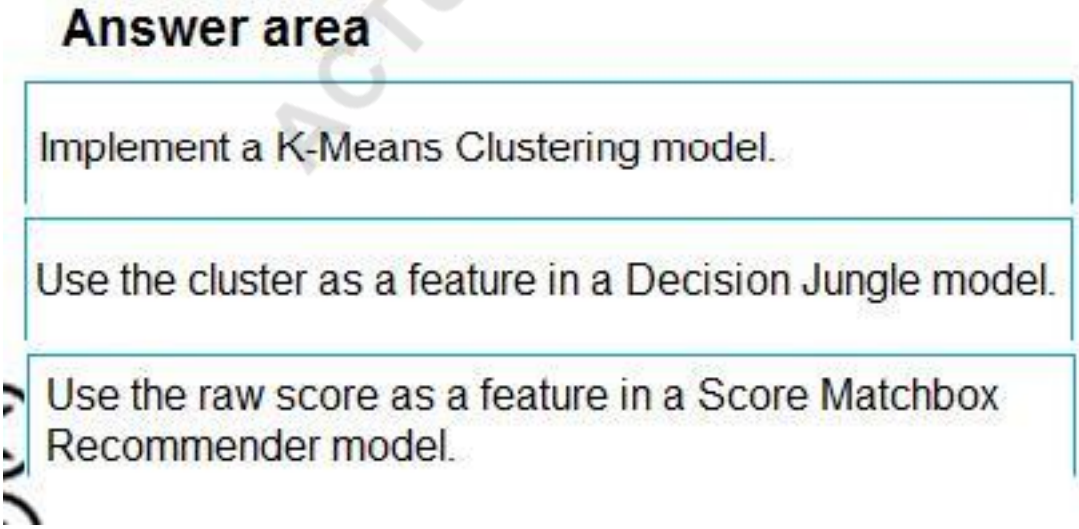
NO.1 You need to define a modeling strategy for ad response.
Which three actions should you perform in sequence? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.



Answer:



Explanation



Step 1: Implement a K-Means Clustering model
Step 2: Use the cluster as a feature in a Decision jungle model.

Decision jungles are non-parametric models, which can represent non-linear decision boundaries.
Step 3: Use the raw score as a feature in a Score Matchbox Recommender model The goal of creating a recommendation system is to recommend one or more "items" to "users" of the system.
Examples of an item could be a movie, restaurant, book, or song. A user could be a person, group of persons, or other entity with item preferences.
Scenario:
Ad response rated declined.
Ad response models must be trained at the beginning of each event and applied during the sporting event.
Market segmentation models must optimize for similar ad response history.
Ad response models must support non-linear boundaries of features.
References:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/multiclass-decision-jungle
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/score-matchbox-recommende

---

Topic 1, Case Study 1
Overview
You are a data scientist in a company that provides data science for professional sporting events. Models will be global and local market data to meet the following business goals:
* Understand sentiment of mobile device users at sporting events based on audio from crowd reactions.
* Access a user's tendency to respond to an advertisement.
* Customize styles of ads served on mobile devices.
* Use video to detect penalty events.
Current environment
Requirements
* Media used for penalty event detection will be provided by consumer devices. Media may include images and videos captured during the sporting event and snared using social media. The images and videos will have varying sizes and formats.
* The data available for model building comprises of seven years of sporting event media. The sporting event media includes: recorded videos, transcripts of radio commentary, and logs from related social media feeds feeds captured during the sporting events.
* Crowd sentiment will include audio recordings submitted by event attendees in both mono and stereo Formats.
Advertisements
* Ad response models must be trained at the beginning of each event and applied during the sporting event.
* Market segmentation nxxlels must optimize for similar ad resporr.r history.
* Sampling must guarantee mutual and collective exclusivity local and global segmentation models that share the same features.
* Local market segmentation models will be applied before determining a user's propensity to respond to an advertisement.
* Data scientists must be able to detect model degradation and decay.
* Ad response models must support non linear boundaries features.

* The ad propensity model uses a cut threshold is 0.45 and retrains occur if weighted Kappa deviates from 0.1
+/-5%.
* The ad propensity model uses cost factors shown in the following diagram:

		Actual	
		1	0
Predicted	0	1	2
	1	2	1

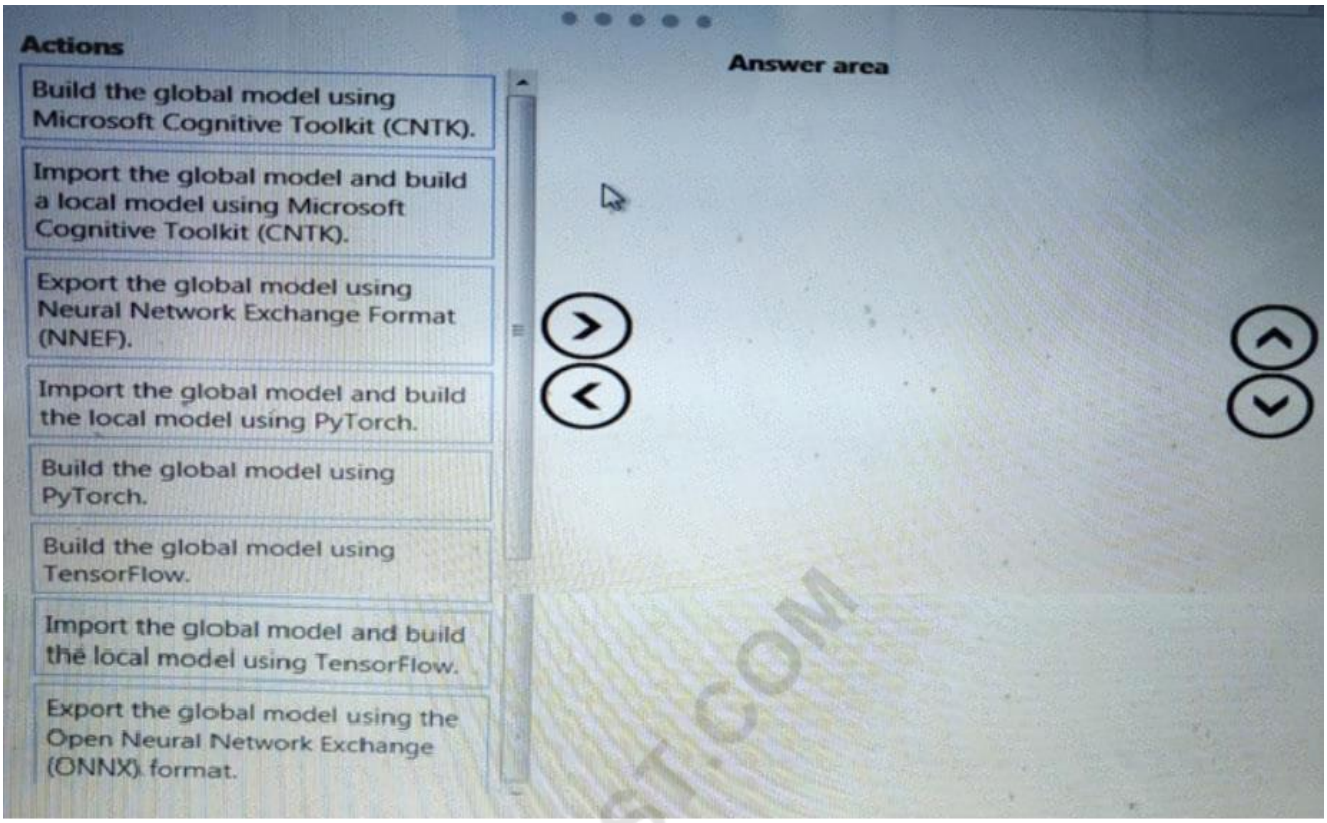
The ad propensity model uses proposed cost factors shown in the following diagram:

		Actual	
		1	0
Predicted	0	1	5
	1	5	1

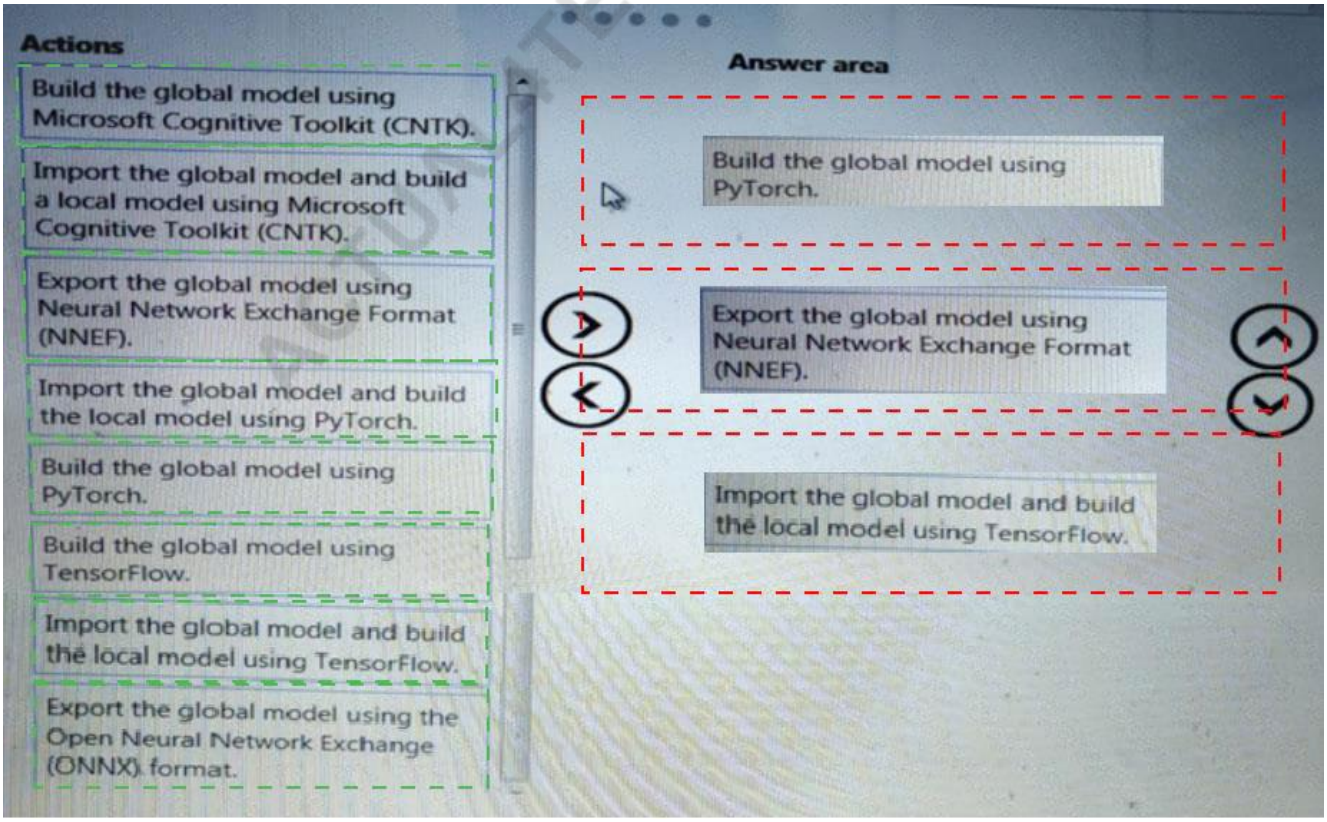
Performance curves of current and proposed cost factor scenarios are shown in the following diagram:

Penalty detection and sentiment
Findings
* Data scientists must build an intelligent solution by using multiple machine learning models for penalty event detection.
* Data scientists must build notebooks in a local environment using automatic feature engineering and model building in machine learning pipelines.
* Notebooks must be deployed to retrain by using Spark instances with dynamic worker allocation
* Notebooks must execute with the same code on new Spark instances to recode only the source of the data.
* Global penalty detection models must be trained by using dynamic runtime graph computation during training.
* Local penalty detection models must be written by using BrainScript.
* Experiments for local crowd sentiment models must combine local penalty detection data.
* Crowd sentiment models must identify known sounds such as cheers and known catch phrases. Individual crowd sentiment models will detect similar sounds.
* All shared features for local models are continuous variables.
* Shared features must use double precision. Subsequent layers must have aggregate running mean and standard deviation metrics Available.
segments
During the initial weeks in production, the following was observed:
* Ad response rates declined.
* Drops were not consistent across ad styles.
* The distribution of features across training and production data are not consistent.
Analysis shows that of the 100 numeric features on user location and behavior, the 47 features that come from location sources are being used as raw features. A suggested experiment to remedy the bias and variance issue is to engineer 10 linearly uncorrected features.
Penalty detection and sentiment
* Initial data discovery shows a wide range of densities of target states in training data used for crowd sentiment models.
* All penalty detection models show inference phases using a Stochastic Gradient Descent (SGD) are running too stow.
* Audio samples show that the length of a catch phrase varies between 25%-47%, depending on region.
* The performance of the global penalty detection models show lower variance but higher bias when comparing training and validation sets. Before implementing any feature changes, you must confirm the bias and variance using all training and validation cases.

NO.2 You need to define a process for penalty event detection.
Which three 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:



NO.3 You need to use the Python language to build a sampling strategy for the global penalty detection models.
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.

import pytorch as deeplearninglib
import tensorflow as deeplearninglib
import cntk as deeplearninglib

train_smapler = deeplearminglib.DistributedSampler.(penalty_video_dataset)
train_sampler = deeplearminglib.log_uniform_candidate_sampler.(penalty_video_dataset)
train_sampler = deeplearninglib.WeightedRandomSampler.(penalty_video_dataset)
train_sampler = deeplearninglib.all_candidate_sampler.(penalty_video_dataset)

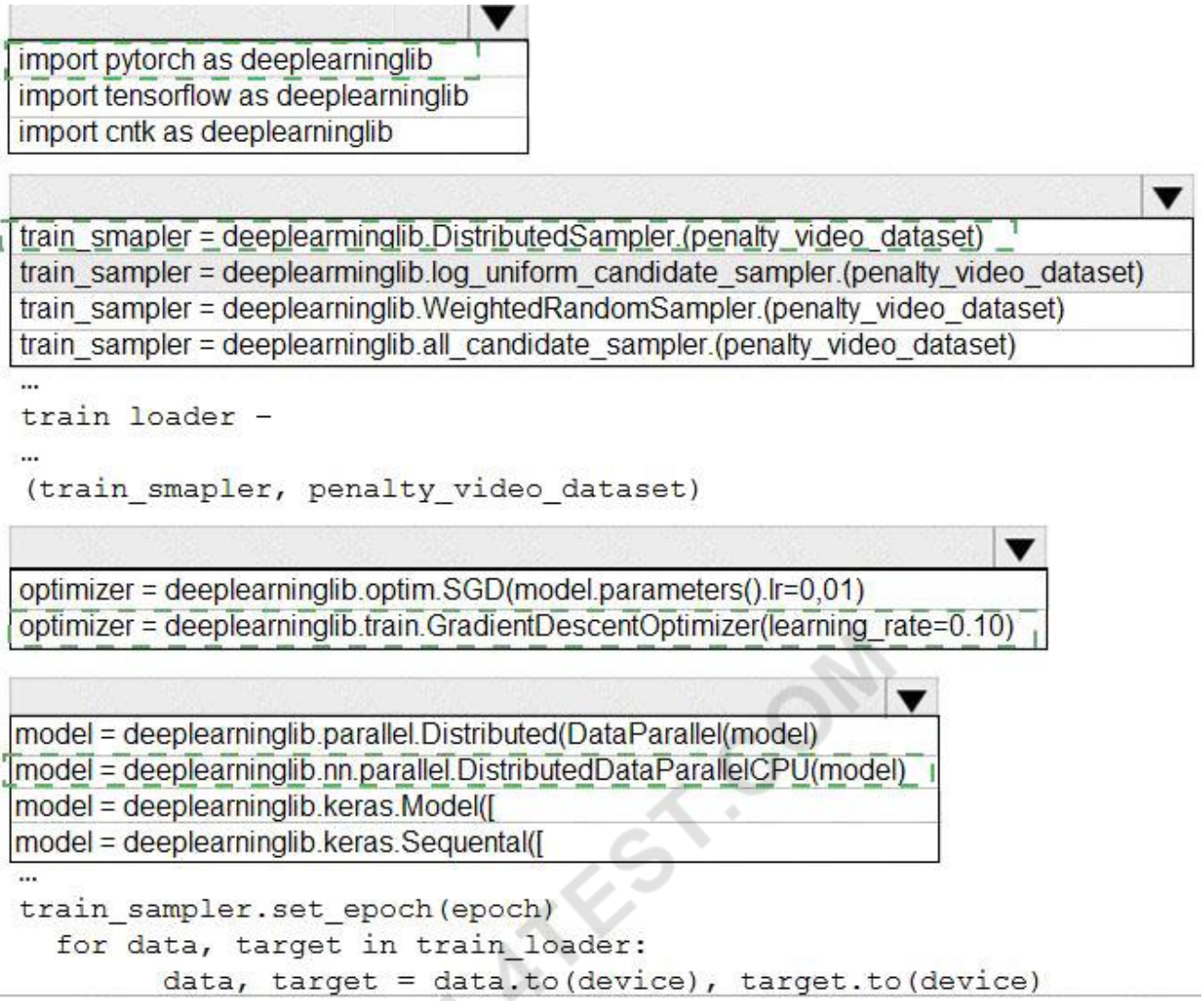
```
...
train loader -
...
(train_smapler, penalty_video_dataset)
```

optimizer = deeplearninglib.optim.SGD(model.parameters().lr=0,01)
optimizer = deeplearninglib.train.GradientDescentOptimizer(learning_rate=0.10)

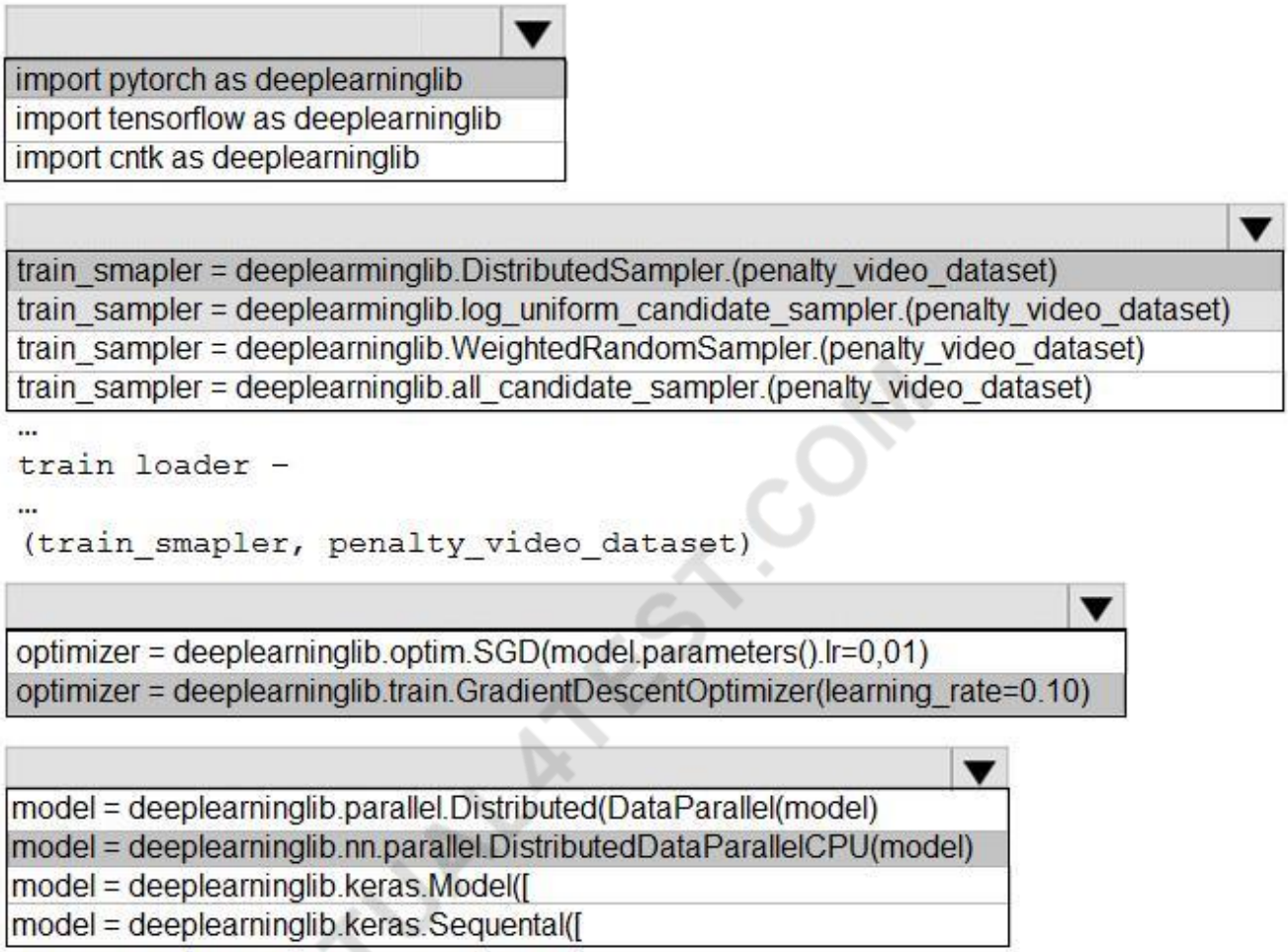
model = deeplearninglib.parallel.Distributed(DataParallel(model)
model = deeplearninglib.nn.parallel.DistributedDataParallelCPU(model)
model = deeplearninglib.keras.Model([
model = deeplearninglib.keras.Sequental([

```
...
train_sampler.set_epoch(epoch)
  for data, target in train_loader:
        data, target = data.to(device), target.to(device)
```

Answer:

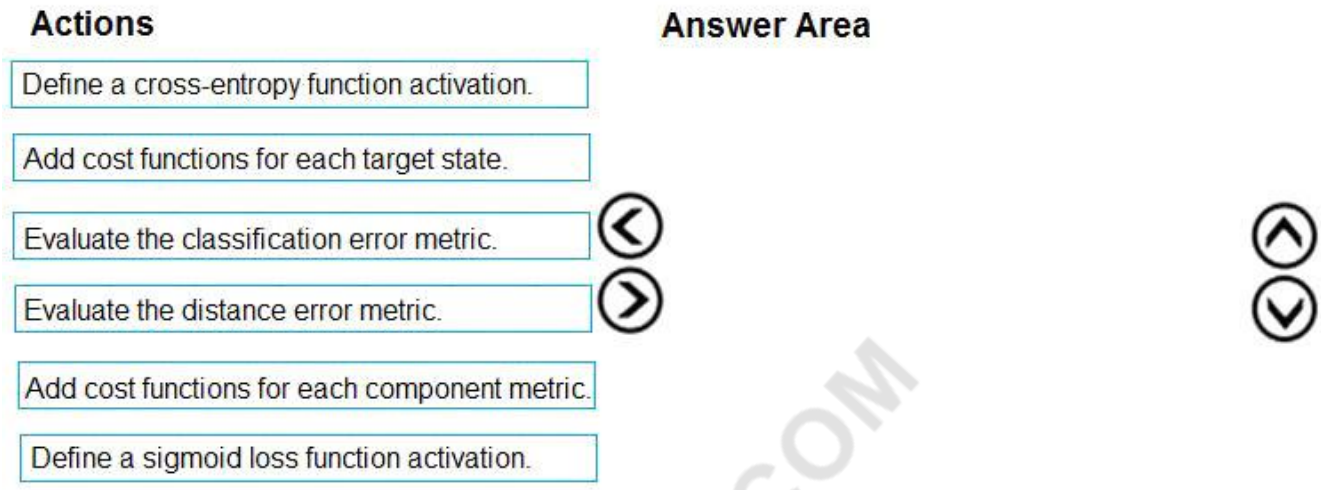


Explanation

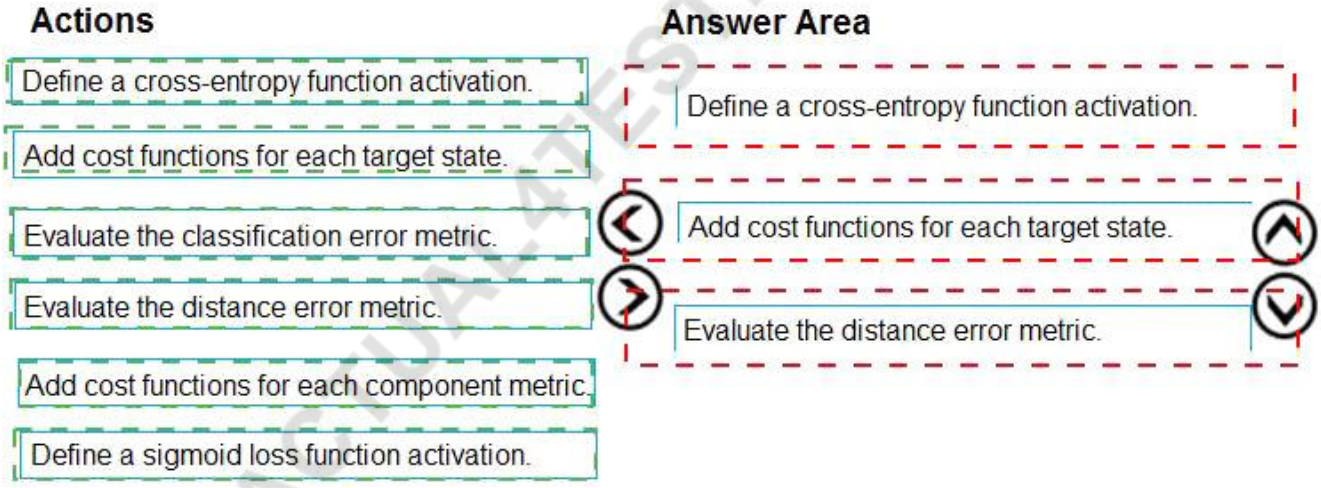


Box 1: import pytorch as deeplearninglib
Box 2: ..DistributedSampler(Sampler)..
DistributedSampler(Sampler):
Sampler that restricts data loading to a subset of the dataset.
It is especially useful in conjunction with class:`torch.nn.parallel.DistributedDataParallel`. In such case, each process can pass a DistributedSampler instance as a DataLoader sampler, and load a subset of the original dataset that is exclusive to it.
Scenario: Sampling must guarantee mutual and collective exclusively between local and global segmentation models that share the same features.
Box 3: optimizer = deeplearninglib.train. GradientDescentOptimizer(learning_rate=0.10)

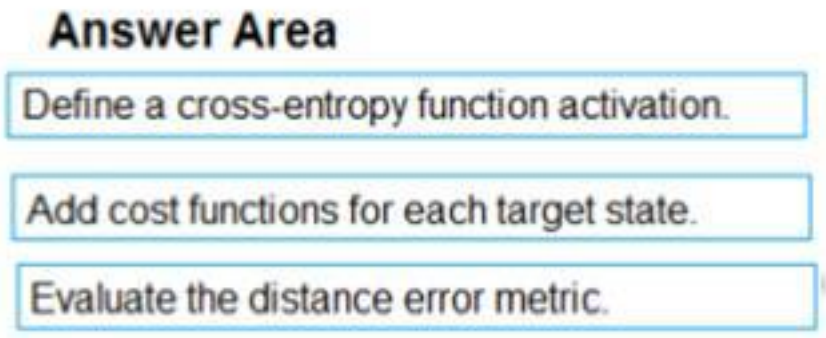
NO.4 You need to define an evaluation strategy for the crowd sentiment models.
Which three actions should you perform in sequence? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.



Answer:



Explanation



Step 1: Define a cross-entropy function activation
When using a neural network to perform classification and prediction, it is usually better to use cross-entropy error than classification error, and somewhat better to use cross-entropy error than mean squared error to evaluate the quality of the neural network.
Step 2: Add cost functions for each target state.
Step 3: Evaluated the distance error metric.
References:
https://www.analyticsvidhya.com/blog/2018/04/fundamentals-deep-learning-regularization-techniques/

NO.5 You need to implement a feature engineering strategy for the crowd sentiment local models.
What should you do?

A. Apply an analysis of variance (ANOVA).

B. Apply a Pearson correlation coefficient.

C. Apply a Spearman correlation coefficient.

D. Apply a linear discriminant analysis.

Answer: D

Explanation

The linear discriminant analysis method works only on continuous variables, not categorical or ordinal variables.

Linear discriminant analysis is similar to analysis of variance (ANOVA) in that it works by comparing the means of the variables.

Scenario:

Data scientists must build notebooks in a local environment using automatic feature engineering and model building in machine learning pipelines.

Experiments for local crowd sentiment models must combine local penalty detection data.

All shared features for local models are continuous variables.

NO.6 You need to select an environment that will meet the business and data requirements.
Which environment should you use?

A. Azure HDInsight with Spark MLlib

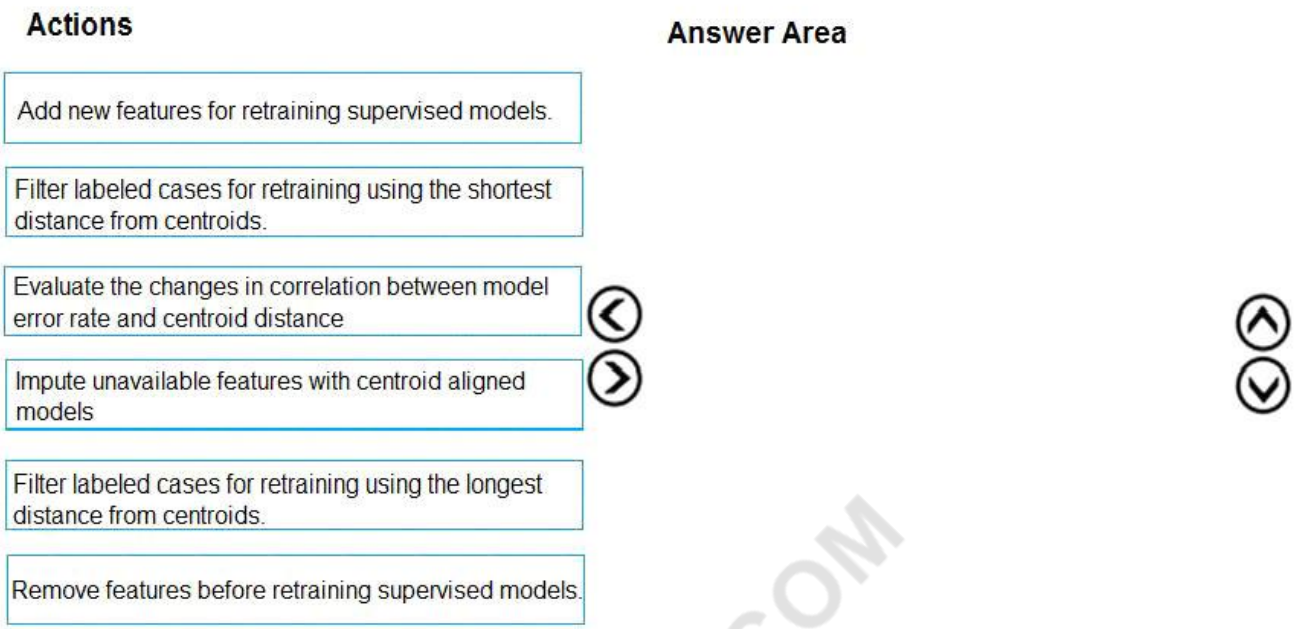
B. Azure Cognitive Services

C. Azure Machine Learning Studio

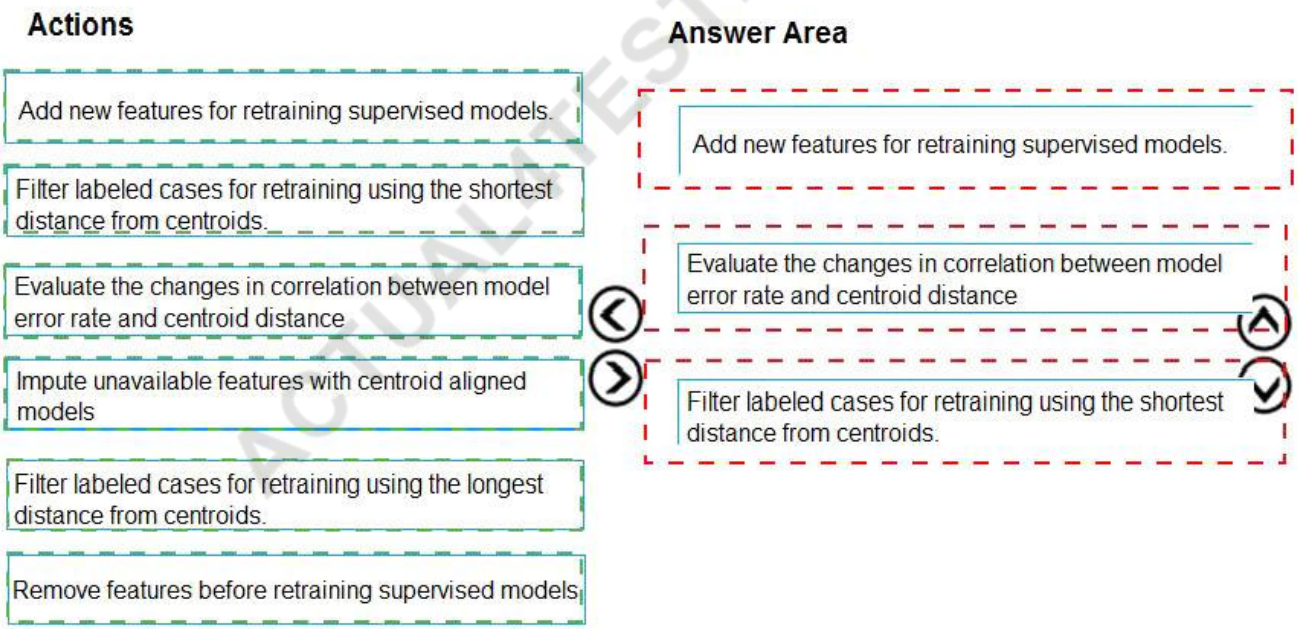
D. Microsoft Machine Learning Server

Answer: D

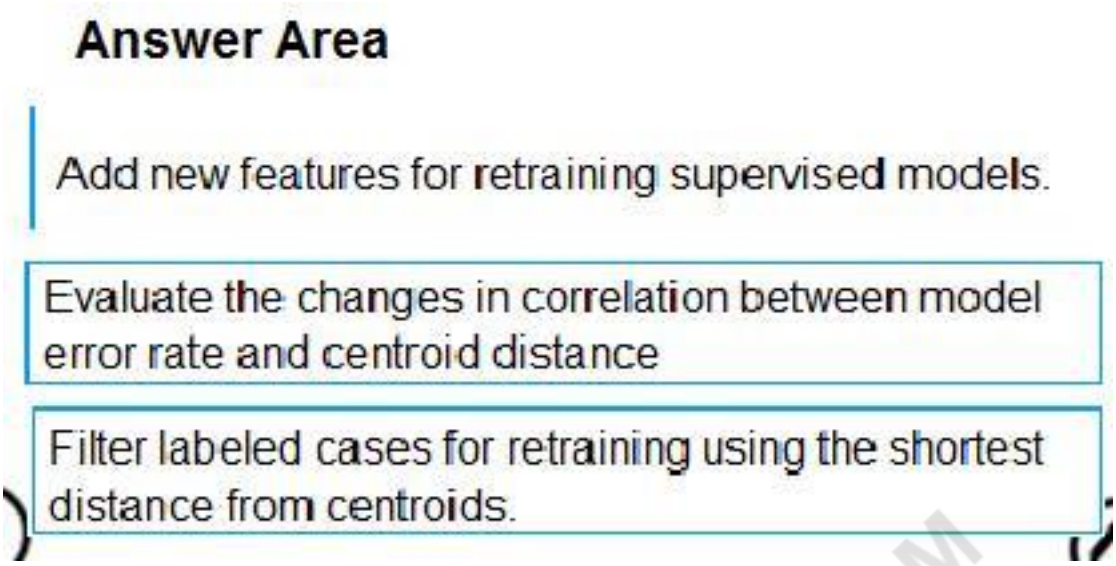
NO.7 You need to define an evaluation strategy for the crowd sentiment models.
Which three actions should you perform in sequence? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.



Answer:

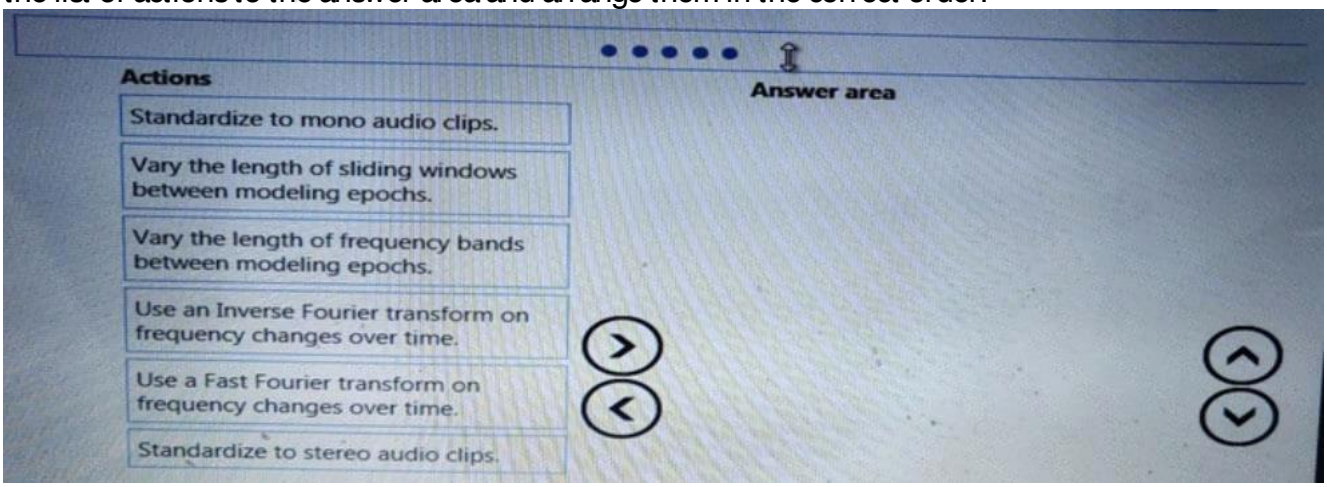


Explanation

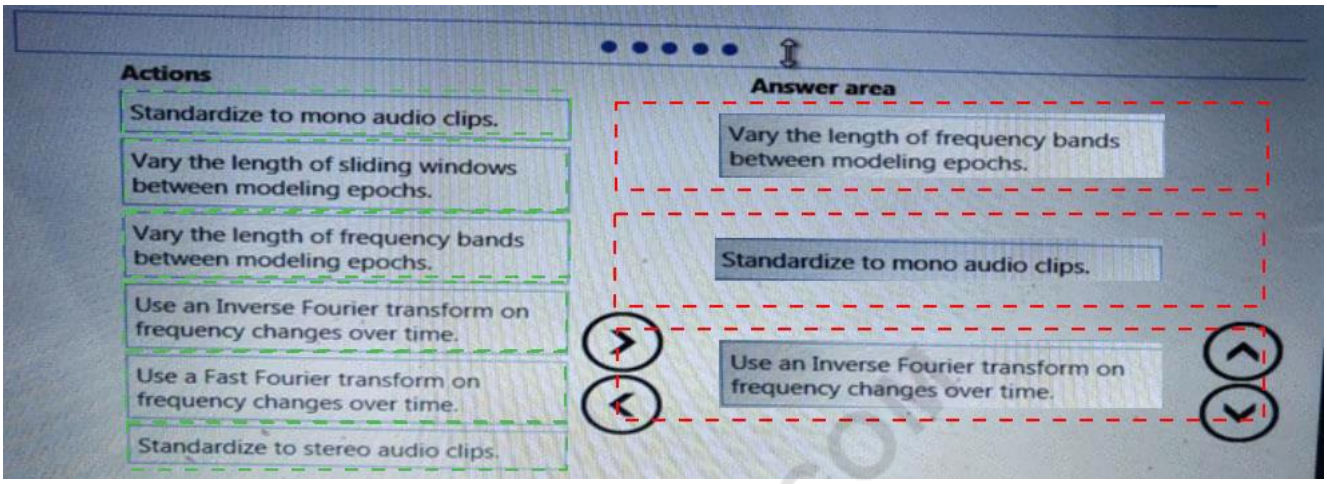


Scenario:
Experiments for local crowd sentiment models must combine local penalty detection data.
Crowd sentiment models must identify known sounds such as cheers and known catch phrases.
Individual crowd sentiment models will detect similar sounds.
Note: Evaluate the changed in correlation between model error rate and centroid distance In machine learning, a nearest centroid classifier or nearest prototype classifier is a classification model that assigns to observations the label of the class of training samples whose mean (centroid) is closest to the observation.
References:
https://en.wikipedia.org/wiki/Nearest_centroid_classifier
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/sweep-clustering

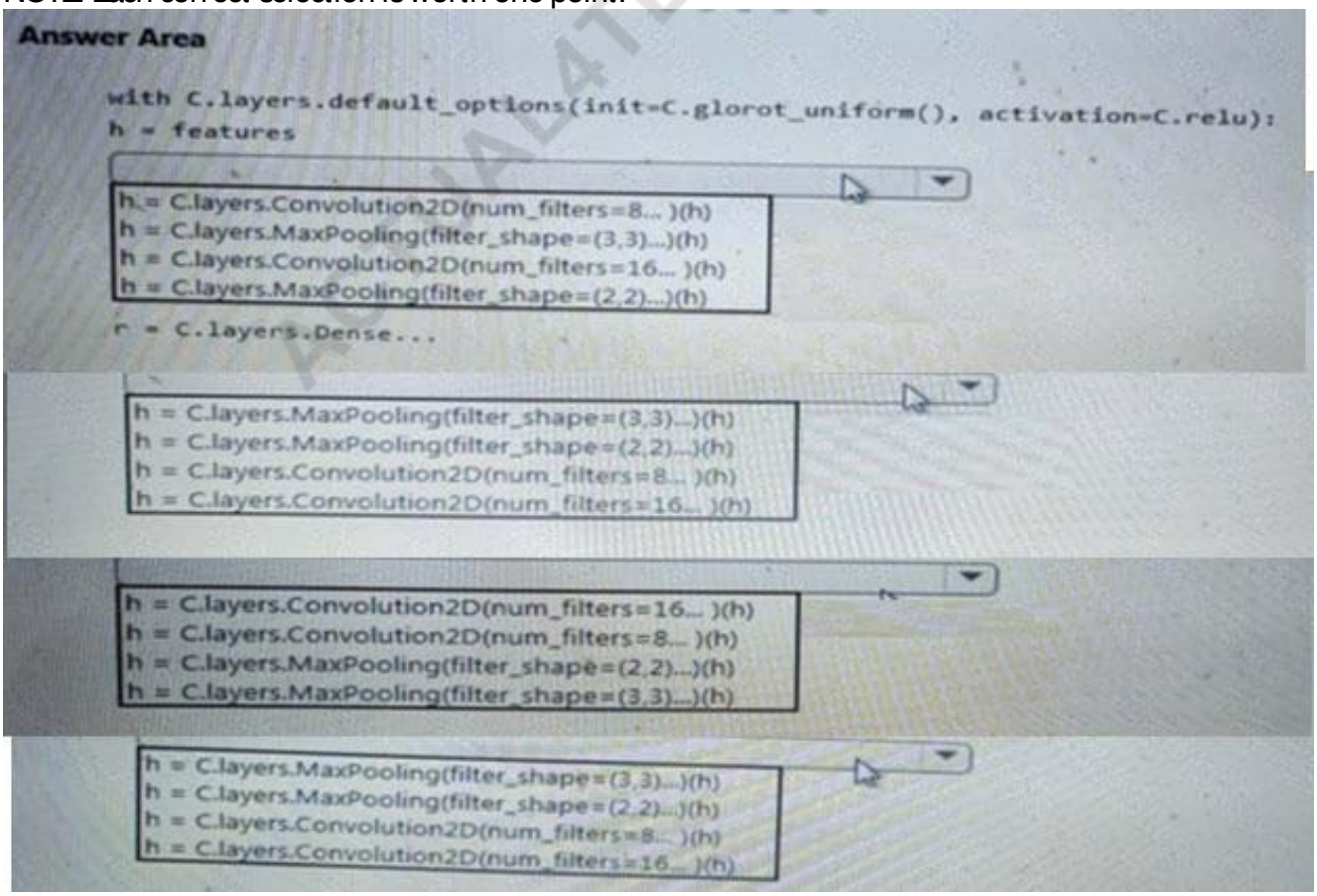
NO.8 You need to define a process for penalty event detection.
Which three actions should you perform in sequence? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.



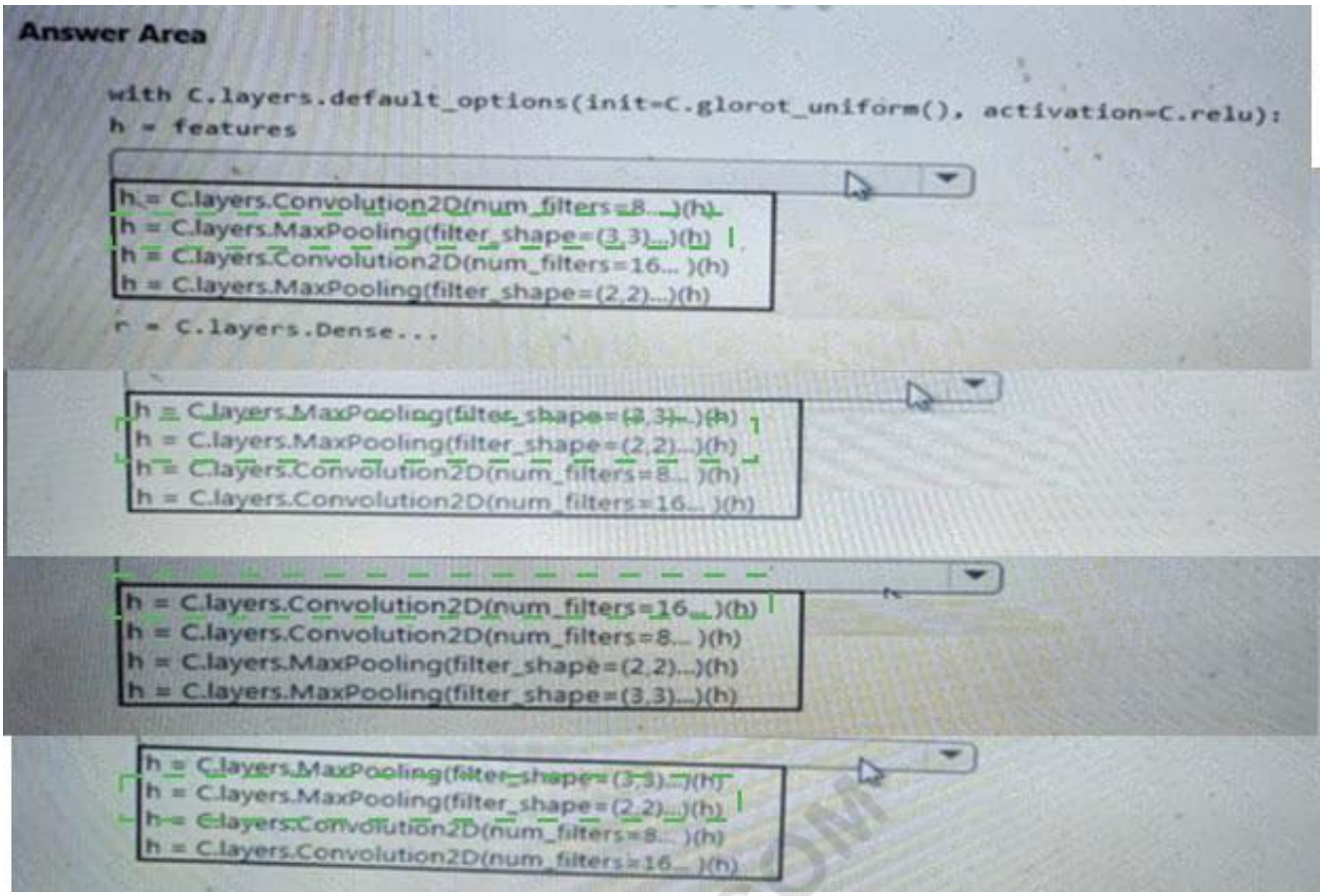
Answer:



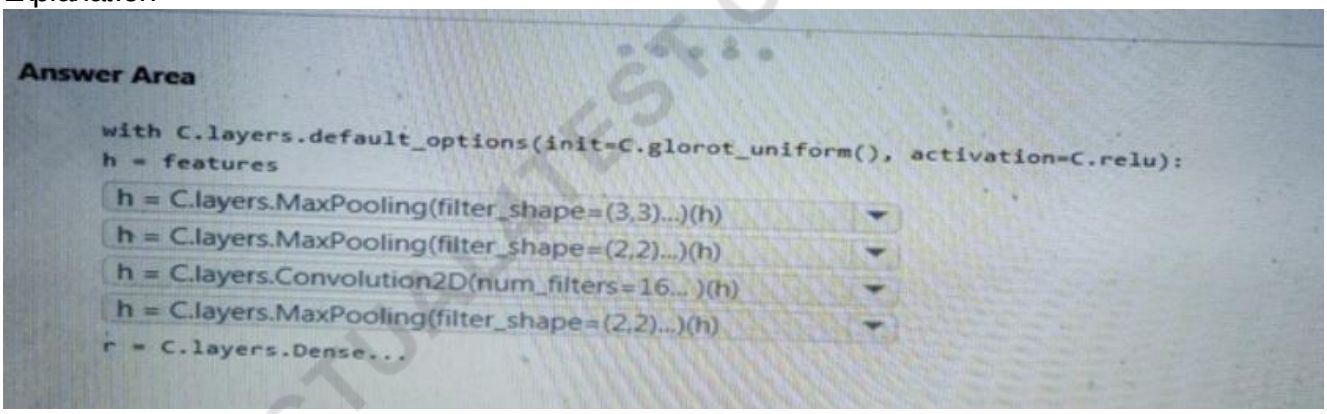
NO.9 You need to build a feature extraction strategy for the local models.
How should you complete the code segment? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.



Answer:



Explanation



NO.10 You need to implement a scaling strategy for the local penalty detection data.
Which normalization type should you use?

A. Streaming

B. Weight

C. Batch

D. Cosine

Answer: C

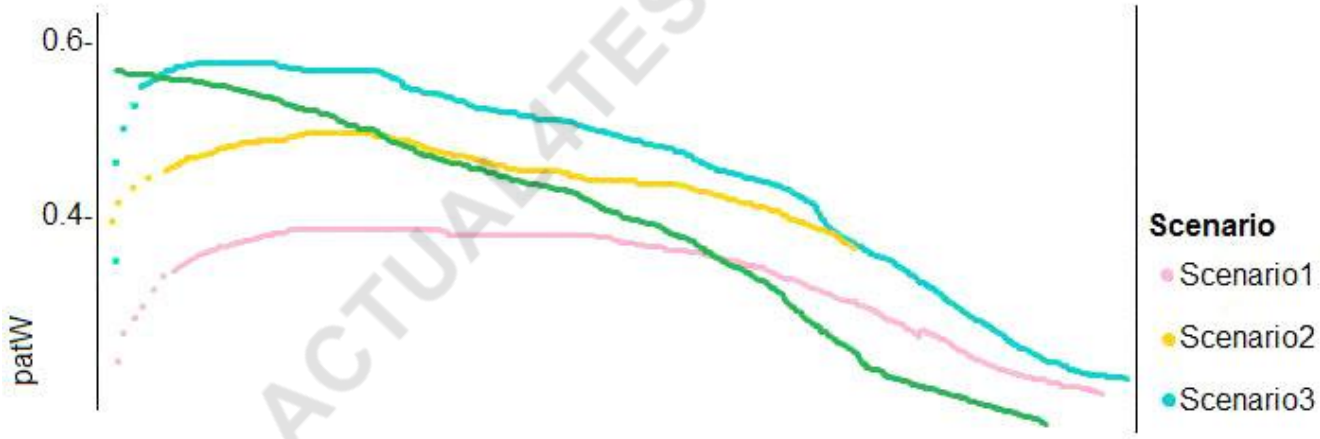
Explanation

Post batch normalization statistics (PBN) is the Microsoft Cognitive Toolkit (CNTK) version of how to evaluate the population mean and variance of Batch Normalization which could be used in inference Original Paper.

In CNTK, custom networks are defined using the BrainScriptNetworkBuilder and described in the

CNTK network description language "BrainScript." Scenario:
Local penalty detection models must be written by using BrainScript.
References:
https://docs.microsoft.com/en-us/cognitive-toolkit/post-batch-normalization-statistics

NO.11 You need to implement a new cost factor scenario for the ad response models as illustrated in the performance curve exhibit.
Which technique should you use?
A. Set the threshold to 0.5 and retrain if weighted Kappa deviates +/- 5% from 0.45.
B. Set the threshold to 0.05 and retrain if weighted Kappa deviates +/- 5% from 0.5.
C. Set the threshold to 0.2 and retrain if weighted Kappa deviates +/- 5% from 0.6.
D. Set the threshold to 0.75 and retrain if weighted Kappa deviates +/- 5% from 0.15.
Answer: A
Explanation
Scenario:
Performance curves of current and proposed cost factor scenarios are shown in the following diagram:



The ad propensity model uses a cut threshold is 0.45 and retrains occur if weighted Kappa deviated from 0.1
+/- 5%.

NO.12 You need to resolve the local machine learning pipeline performance issue. What should you do?
A. Increase Graphic Processing Units (GPUs).
B. Increase the learning rate.
C. Increase the training iterations,
D. Increase Central Processing Units (CPUs).
Answer: A

NO.13 You need to implement a model development strategy to determine a user's tendency to respond to an ad.
Which technique should you use?
A. Use a Relative Expression Split module to partition the data based on centroid distance.

B. Use a Relative Expression Split module to partition the data based on distance travelled to the event.

C. Use a Split Rows module to partition the data based on distance travelled to the event.

D. Use a Split Rows module to partition the data based on centroid distance.

Answer: A

Explanation

Split Data partitions the rows of a dataset into two distinct sets.

The Relative Expression Split option in the Split Data module of Azure Machine Learning Studio is helpful when you need to divide a dataset into training and testing datasets using a numerical expression.

Relative Expression Split: Use this option whenever you want to apply a condition to a number column. The number could be a date/time field, a column containing age or dollar amounts, or even a percentage. For example, you might want to divide your data set depending on the cost of the items, group people by age ranges, or separate data by a calendar date.

Scenario:

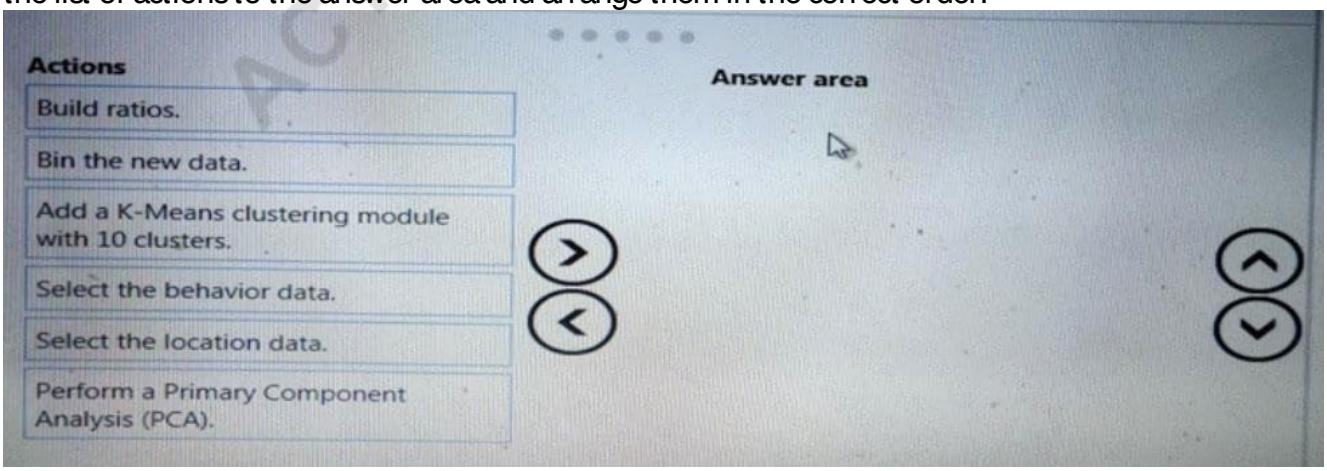
Local market segmentation models will be applied before determining a user's propensity to respond to an advertisement.

The distribution of features across training and production data are not consistent References:

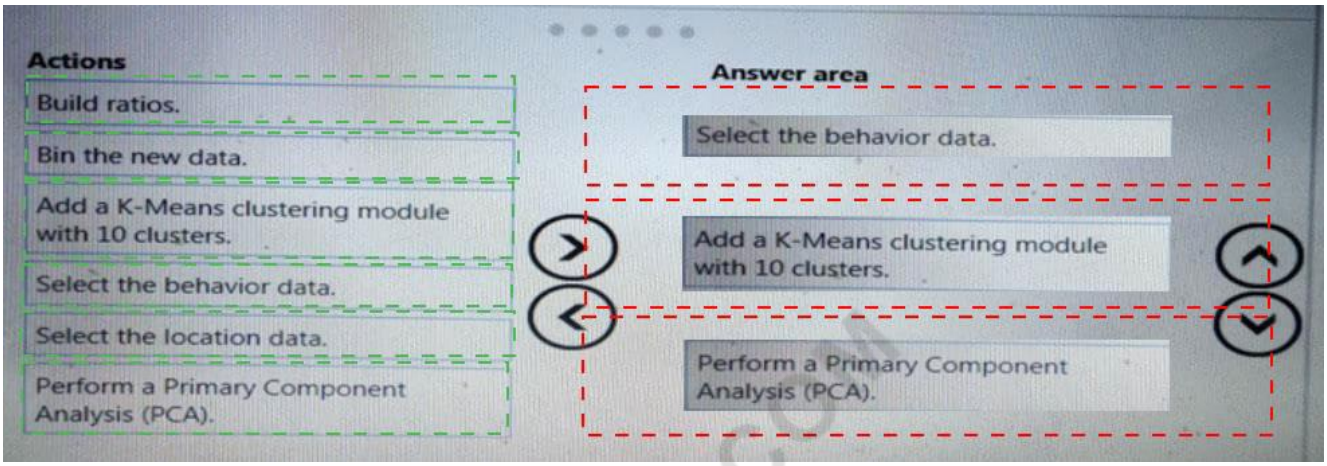
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/split-data

NO.14 You need to modify the inputs for the global penalty event model to address the bias and variance issue.

Which three 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:



Topic 2, Case Study 2

Case study

Overview

You are a data scientist for Fabrikam Residences, a company specializing in quality private and commercial property in the United States. Fabrikam Residences is considering expanding into Europe and has asked you to investigate prices for private residences in major European cities. You use Azure Machine Learning Studio to measure the median value of properties. You produce a regression model to predict property prices by using the Linear Regression and Bayesian Linear Regression modules.

Datasets

There are two datasets in CSV format that contain property details for two cities, London and Paris, with the following columns:

Column heading	Description
CapitaCrimeRate	per capita crime rate by town
Zoned	proportion of residential land zoned for lots over 25.000 square feet
NonRetailAcres	proportion of retail business acres per town
NextToRiver	proximity of the property to the river
NitrogenOxideConcentration	nitric oxides concentration (parts per 10 million)
AvgRoomsPerHouse	average number of rooms per dwelling
Age	proportion of owner-occupied units built prior to 1940
DistanceToEmploymentCenter	weighted distances to employment centers
AccessibilityToHighway	index of accessibility to radial highways to a value of two decimal places
Tax	full value property tax rate per $10,000
PupilTeacherRatio	pupil to teacher ratio by town
ProfessionalClass	professional class percentage
LowerStatus	percentage lower status of the population
MedianValue	median value of owner-occupied homes in $1000s

The two datasets have been added to Azure Machine Learning Studio as separate datasets and included as the starting point of the experiment.

Dataset issues

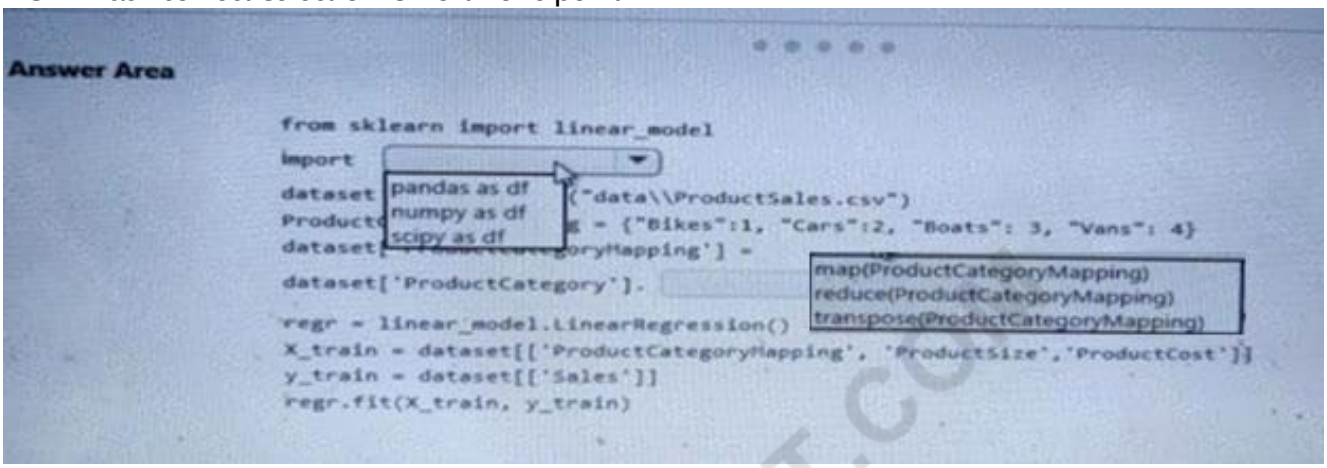
The AccessibilityToHighway column in both datasets contains missing values. The missing data must be replaced with new data so that it is modeled conditionally using the other variables in the data before filling in the missing values.

Columns in each dataset contain missing and null values. The dataset also contains many outliers. The

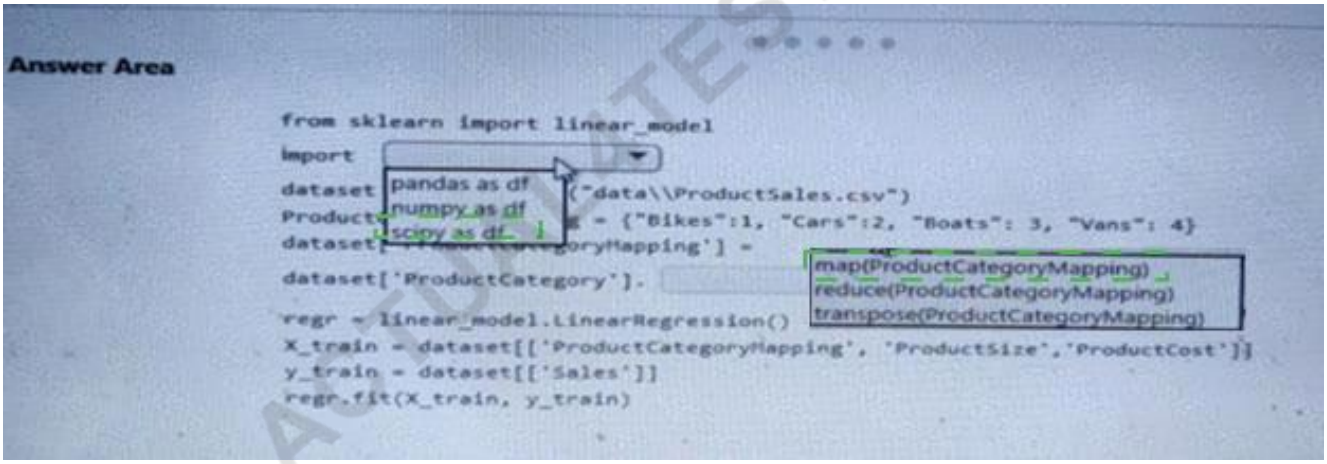
Age column has a high proportion of outliers. You need to remove the rows that have outliers in the Age column.
The MedianValue and AvgRoomsinHouse columns both hold data in numeric format. You need to select a feature selection algorithm to analyze the relationship between the two columns in more detail.
Model fit
The model shows signs of overfitting. You need to produce a more refined regression model that reduces the overfitting.
Experiment requirements
You must set up the experiment to cross-validate the Linear Regression and Bayesian Linear Regression modules to evaluate performance.
In each case, the predictor of the dataset is the column named MedianValue. An initial investigation showed that the datasets are identical in structure apart from the MedianValue column. The smaller Paris dataset contains the MedianValue in text format, whereas the larger London dataset contains the MedianValue in numerical format. You must ensure that the datatype of the MedianValue column of the Paris dataset matches the structure of the London dataset.
You must prioritize the columns of data for predicting the outcome. You must use non-parameters statistics to measure the relationships.
You must use a feature selection algorithm to analyze the relationship between the MedianValue and AvgRoomsinHouse columns.
Model training
Given a trained model and a test dataset, you need to compute the permutation feature importance scores of feature variables. You need to set up the Permutation Feature Importance module to select the correct metric to investigate the model's accuracy and replicate the findings.
You want to configure hyperparameters in the model learning process to speed the learning phase by using hyperparameters. In addition, this configuration should cancel the lowest performing runs at each evaluation interval, thereby directing effort and resources towards models that are more likely to be successful.
You are concerned that the model might not efficiently use compute resources in hyperparameter tuning. You also are concerned that the model might prevent an increase in the overall tuning time. Therefore, you need to implement an early stopping criterion on models that provides savings without terminating promising jobs.
Testing
You must produce multiple partitions of a dataset based on sampling using the Partition and Sample module in Azure Machine Learning Studio. You must create three equal partitions for cross-validation. You must also configure the cross-validation process so that the rows in the test and training datasets are divided evenly by properties that are near each city's main river. The data that identifies that a property is near a river is held in the column named NextToRiver. You want to complete this task before the data goes through the sampling process.
When you train a Linear Regression module using a property dataset that shows data for property prices for a large city, you need to determine the best features to use in a model. You can choose standard metrics provided to measure performance before and after the feature importance process completes. You must ensure that the distribution of the features across multiple training models is consistent.
Data visualization
You need to provide the test results to the Fabrikam Residences team. You create data visualizations

* Vans
* Boats

You are building a regression model using the scikit- learn Python package.
You need to transform the text data to be compatible with the scikit-learn Python package 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:



NO.19 You are performing feature engineering on a dataset.
You must add a feature named CityName and populate the column value with the text London.
You need to add the new feature to the dataset.
Which Azure Machine Learning Studio module should you use?

A. Edit Metadata

B. Preprocess Text

C. Execute Python Script

D. Latent Dirichlet Allocation

Answer: A

Explanation
Typical metadata changes might include marking columns as features.
References:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/edit-metadata

NO.20 You use Azure Machine Learning Studio to build a machine learning experiment.
You need to divide data into two distinct datasets.
Which module should you use?
A. Partition and Sample
B. Assign Data to Clusters
C. Group Data into Bins
D. Test Hypothesis Using t-Test
Answer: A
Explanation
Partition and Sample with the Stratified split option outputs multiple datasets, partitioned using the rules you specified.
References:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/partition-and-sample

NO.21 You are analyzing a raw dataset that requires cleaning.
You must perform transformations and manipulations by using Azure Machine Learning Studio.
You need to identify the correct modules to perform the transformations.
Which modules should you choose? To answer, drag the appropriate modules to the correct scenarios. Each module may be used once, 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

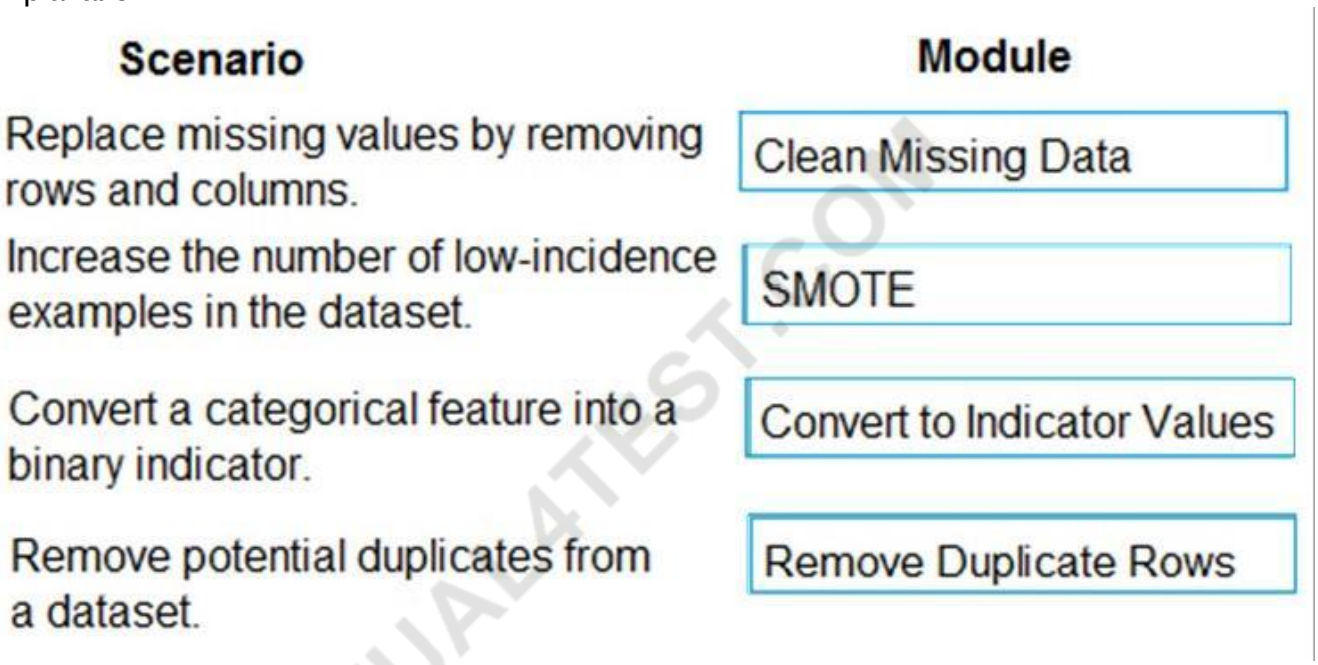
Methods	Scenario	Module
Clean Missing Data	Replace missing values by removing rows and columns.	
SMOTE	Increase the number of low-incidence examples in the dataset.	
Convert to Indicator Values	Convert a categorical feature into a binary indicator.	
Remove Duplicate Rows	Remove potential duplicates from a dataset.	
Threshold Filter		

Answer:

Answer Area

Methods	Scenario	Module
Clean Missing Data	Replace missing values by removing rows and columns.	Clean Missing Data
SMOTE	Increase the number of low-incidence examples in the dataset.	SMOTE
Convert to Indicator Values	Convert a categorical feature into a binary indicator.	Convert to Indicator Values
Remove Duplicate Rows	Remove potential duplicates from a dataset.	Remove Duplicate Rows
Threshold Filter		

Explanation



Box 1: Clean Missing Data
Box 2: SMOTE
Use the SMOTE module in Azure Machine Learning Studio to increase the number of underepresented cases in a dataset used for machine learning. SMOTE is a better way of increasing the number of rare cases than simply duplicating existing cases.
Box 3: Convert to Indicator Values
Use the Convert to Indicator Values module in Azure Machine Learning Studio. The purpose of this module is to convert columns that contain categorical values into a series of binary indicator columns that can more easily be used as features in a machine learning model.
Box 4: Remove Duplicate Rows
References:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/smote
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/convert-to-indicator-values

NO.22 You are a data scientist building a deep convolutional neural network (CNN) for image

classification.
The CNN model you built shows signs of overfitting.
You need to reduce overfitting and converge the model to an optimal fit.
Which two actions should you perform? Each correct answer presents a complete solution.
NOTE: Each correct selection is worth one point.

A. Reduce the amount of training data.
B. Add an additional dense layer with 64 input units
C. Add L1/L2 regularization.
D. Use training data augmentation
E. Add an additional dense layer with 512 input units.

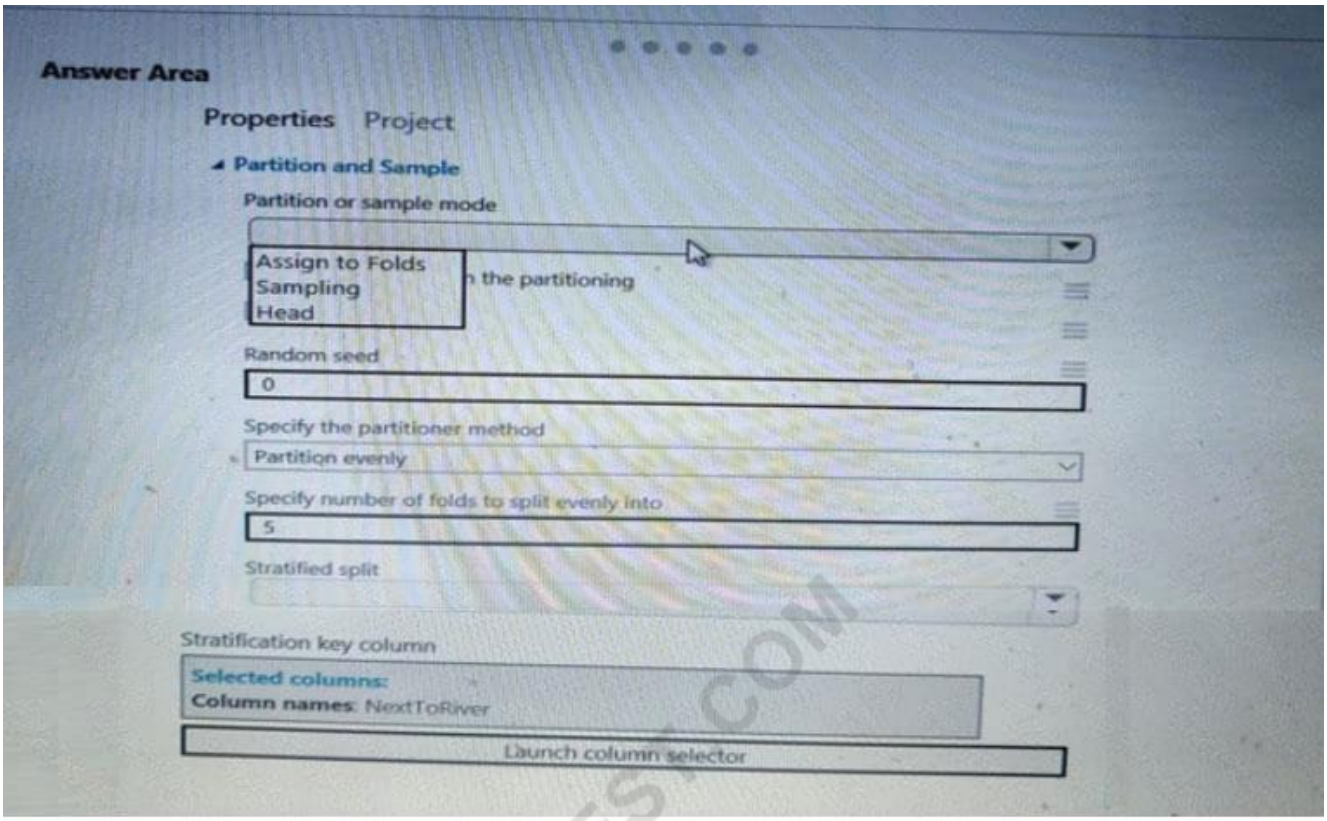
Answer: B E

NO.23 You create a classification model with a dataset that contains 100 samples with Class A and 10,000 samples with Class B The variation of Class B is very high.
You need to resolve imbalances.
Which method should you use?

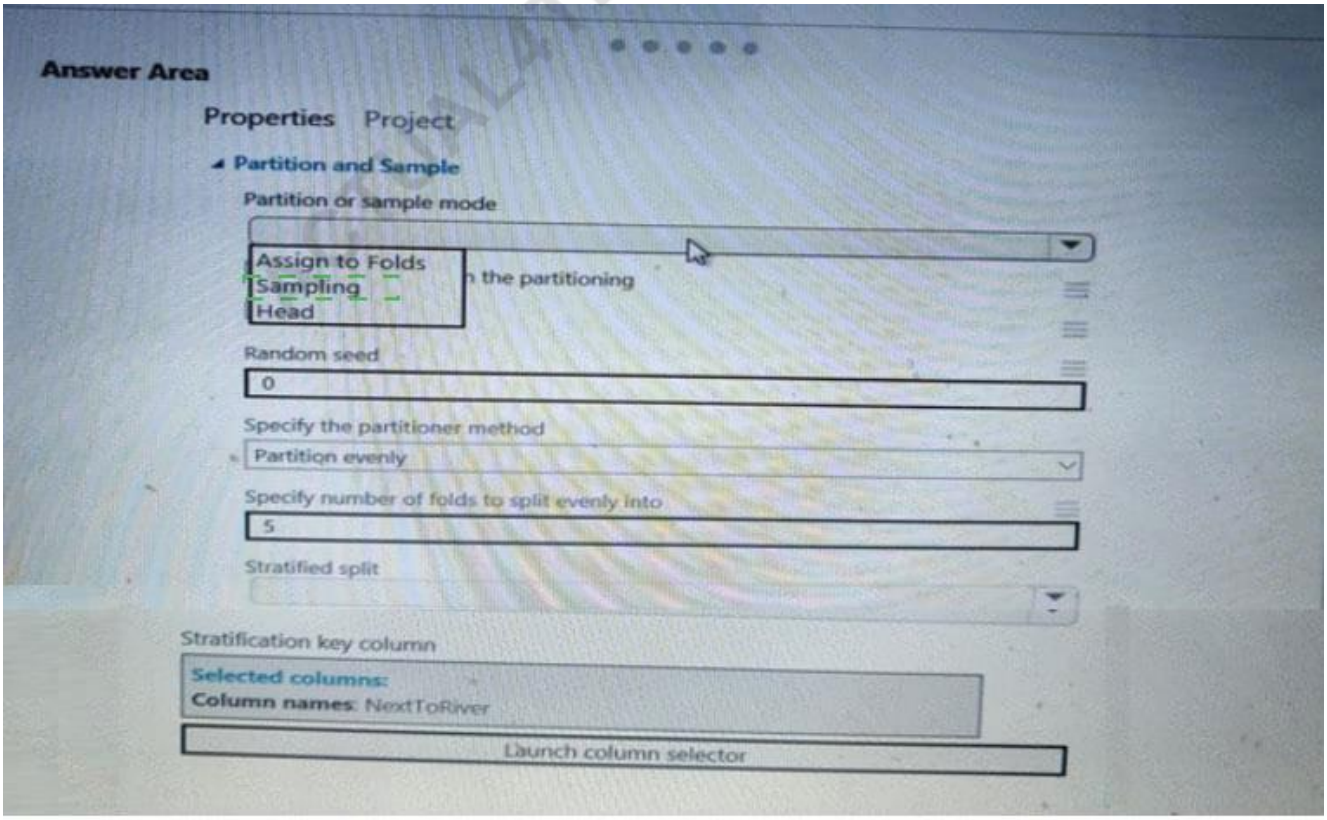
A. Partition and Sample
B. Cluster Centroids
C. Tomek links
D. Synthetic Minority Oversampling Technique (SMOTE)

Answer: D

NO.24 You need to identify the methods for dividing the data according, to the testing requirements.
Which properties should you select? To answer, select the appropriate option-, m the answer area.
NOTE:
Each correct selection is worth one point.



Answer:



NO.25 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 contains 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: Replace each missing value using the Multiple Imputation by Chained Equations (MICE) method.
Does the solution meet the goal?

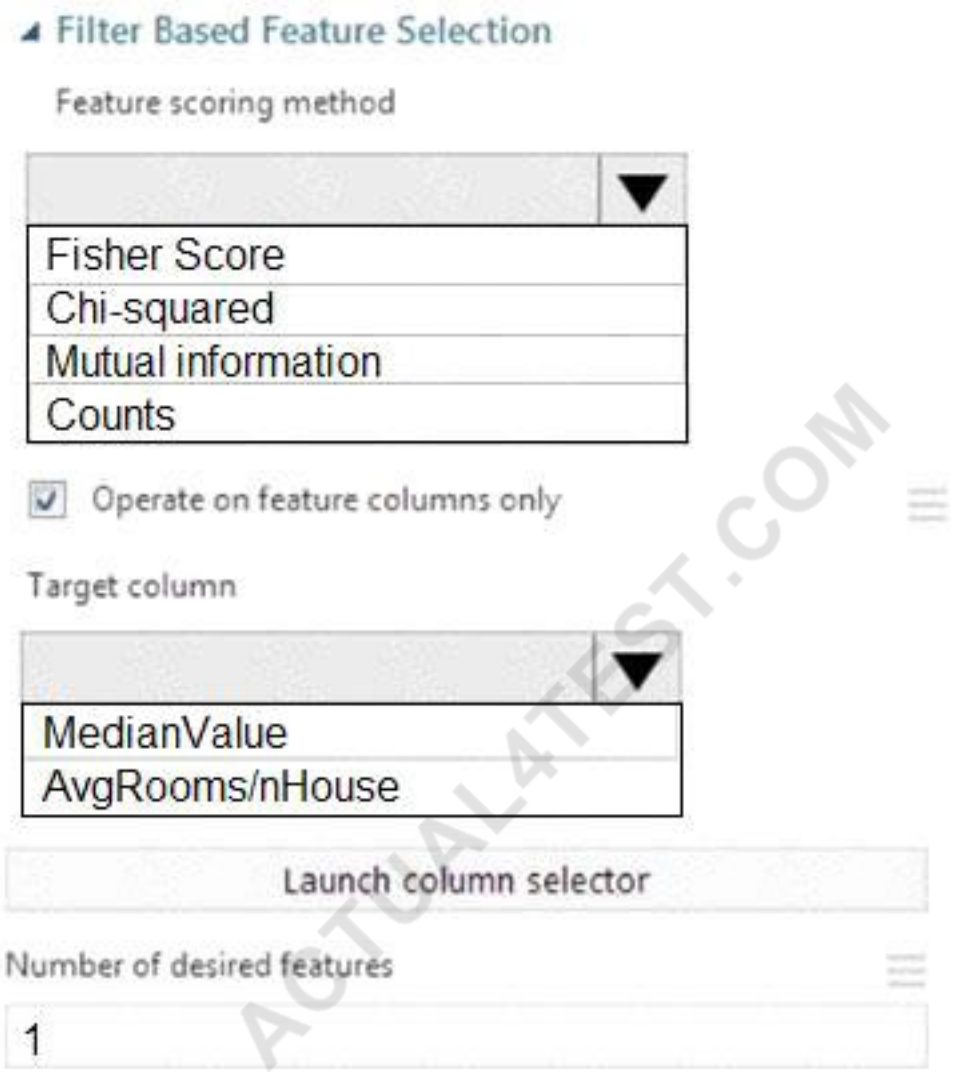
A. Yes

B. NO

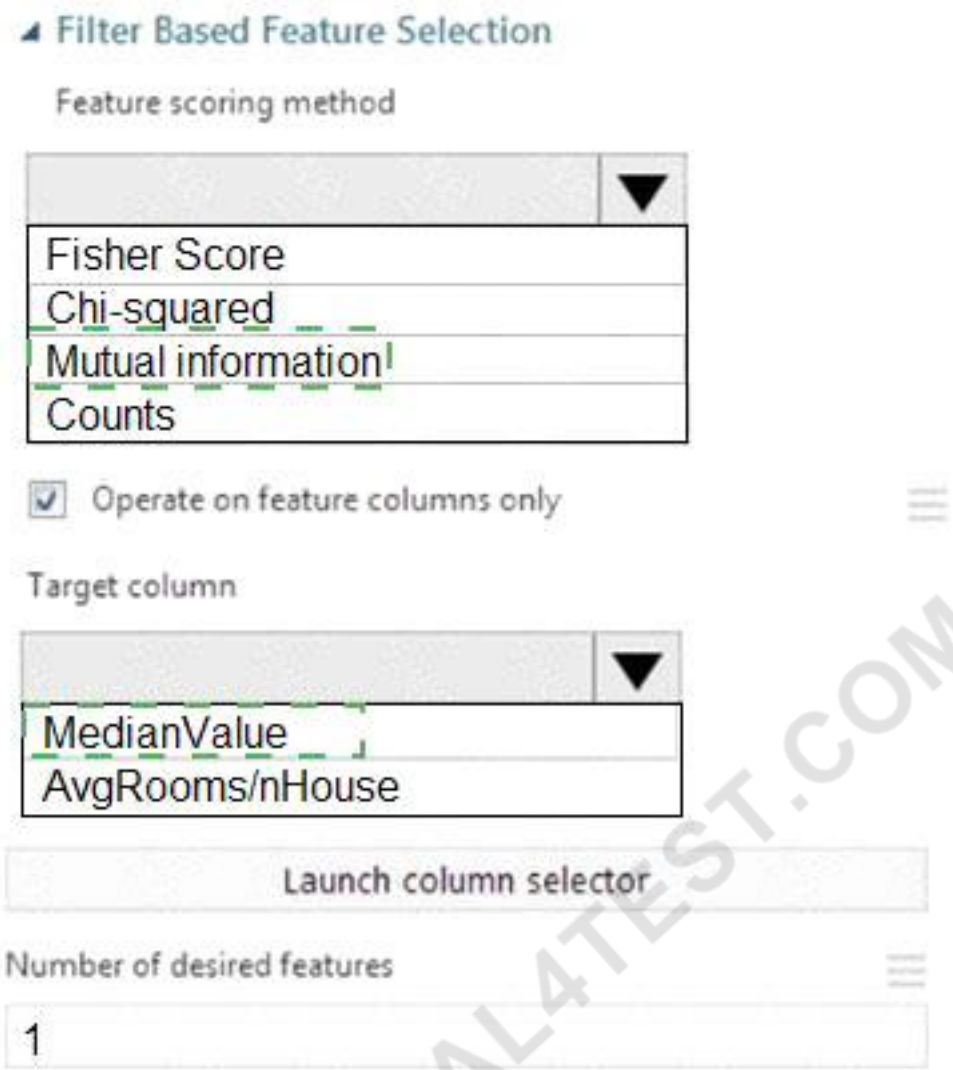
Answer: A

Explanation
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: Multivariate imputation by chained equations (MICE), sometimes called "fully conditional specification" or "sequential regression multiple imputation" has emerged in the statistical literature as one principled method of addressing missing data. Creating multiple imputations, as opposed to single imputations, accounts for the statistical uncertainty in the imputations. In addition, the chained equations approach is very flexible and can handle variables of varying types (e.g., continuous or binary) as well as complexities such as bounds or survey skip patterns.
References:
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3074241/
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/clean-missing-data

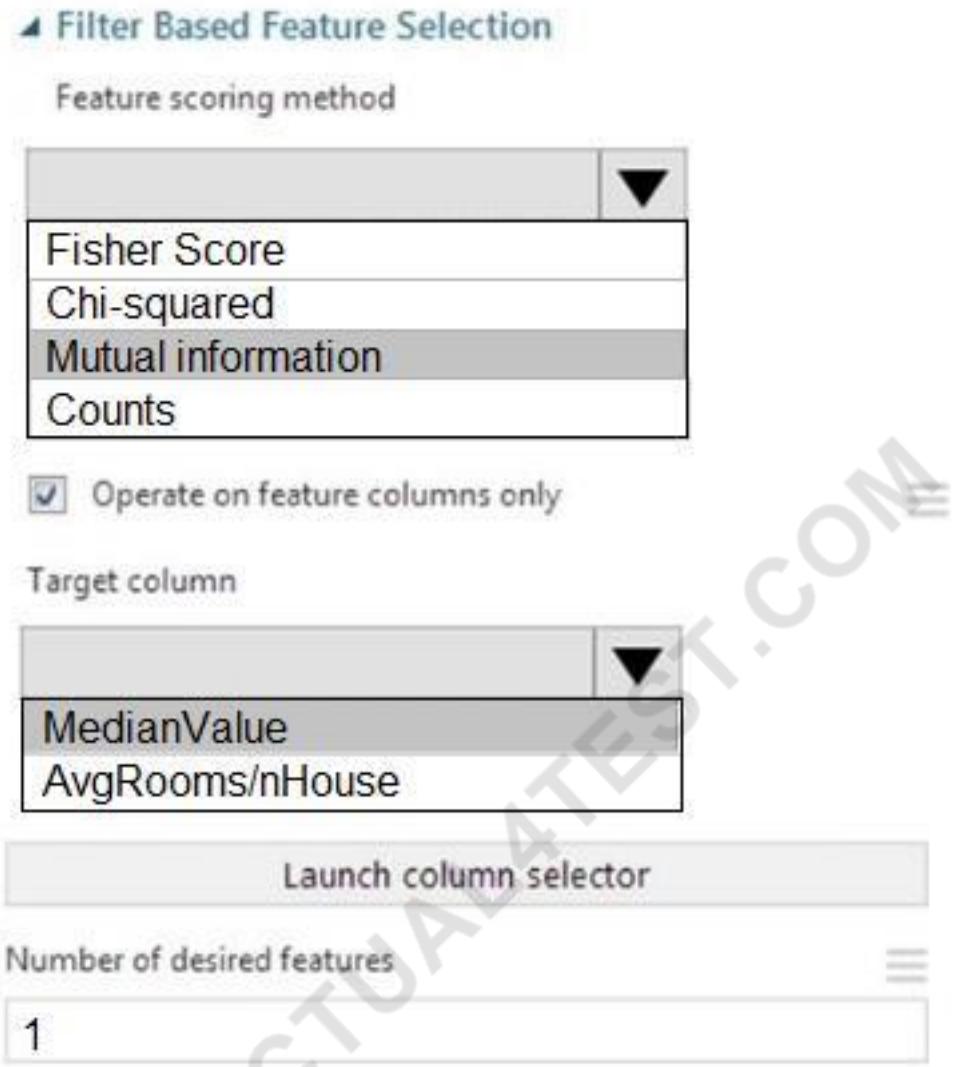
NO.26 You need to configure the Feature Based Feature Selection module based on the experiment requirements and datasets.
How should you configure the module properties? To answer, select the appropriate options in the dialog box in the answer area.
NOTE: Each correct selection is worth one point.



Answer:



Explanation



Box 1: Mutual Information.
The mutual information score is particularly useful in feature selection because it maximizes the mutual information between the joint distribution and target variables in datasets with many dimensions.
Box 2: MedianValue
MedianValue is the feature column, , it is the predictor of the dataset.
Scenario: The MedianValue and AvgRoomsinHouse columns both hold data in numeric format. You need to select a feature selection algorithm to analyze the relationship between the two columns in more detail.
References:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/filter-based-feature-selection

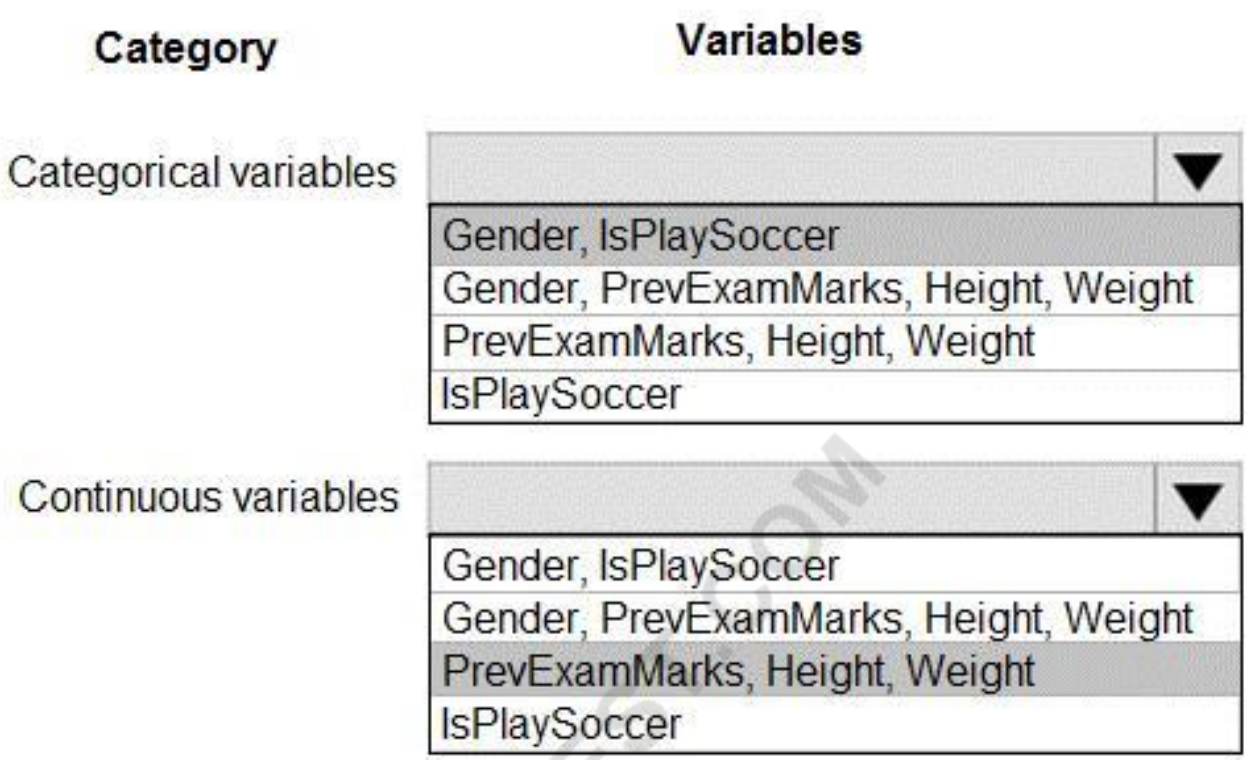
NO.27 You are working on a classification task. You have a dataset indicating whether a student would like to play soccer and associated attributes. The dataset includes the following columns:
You need to classify variables by type.
Which variable should you add to each category? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.

Category	Variables
Categorical variables	Gender, IsPlaySoccer Gender, PrevExamMarks, Height, Weight PrevExamMarks, Height, Weight IsPlaySoccer
Continuous variables	Gender, IsPlaySoccer Gender, PrevExamMarks, Height, Weight PrevExamMarks, Height, Weight IsPlaySoccer

Answer:

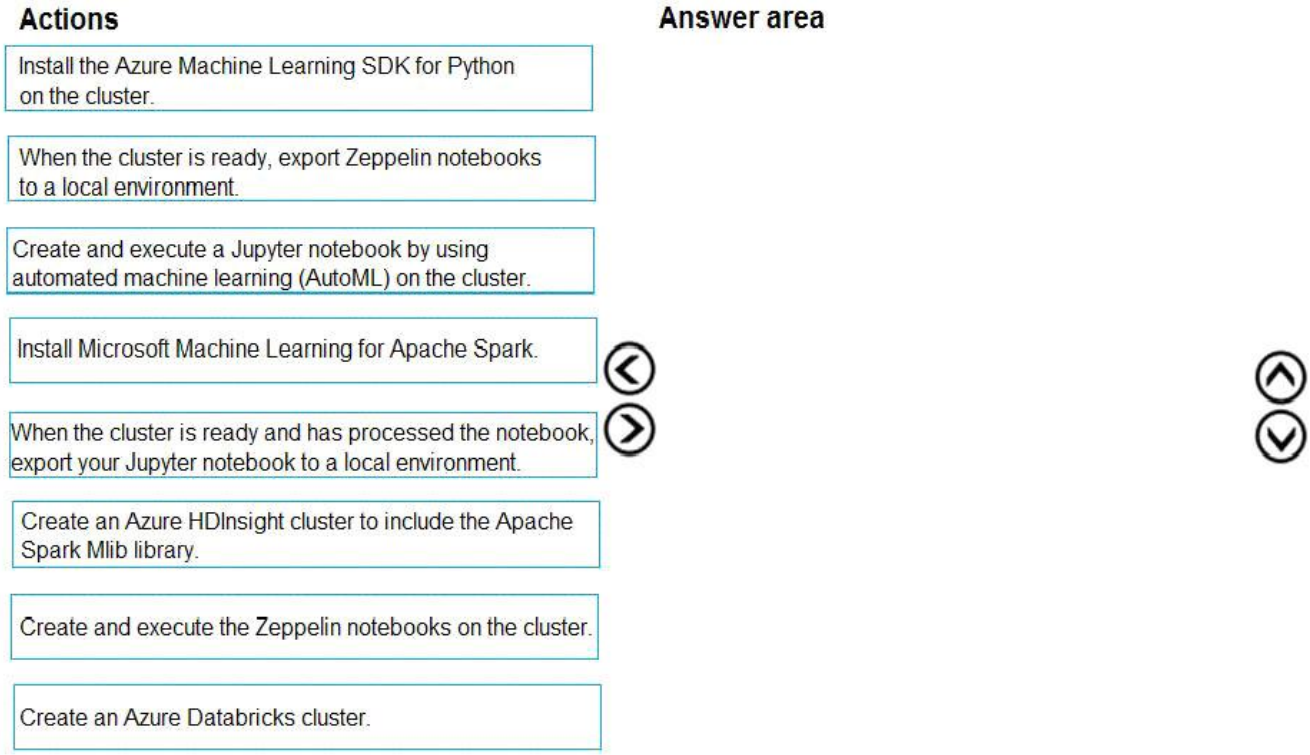
Category	Variables
Categorical variables	Gender, IsPlaySoccer Gender, PrevExamMarks, Height, Weight PrevExamMarks, Height, Weight IsPlaySoccer
Continuous variables	Gender, IsPlaySoccer Gender, PrevExamMarks, Height, Weight PrevExamMarks, Height, Weight IsPlaySoccer

Explanation

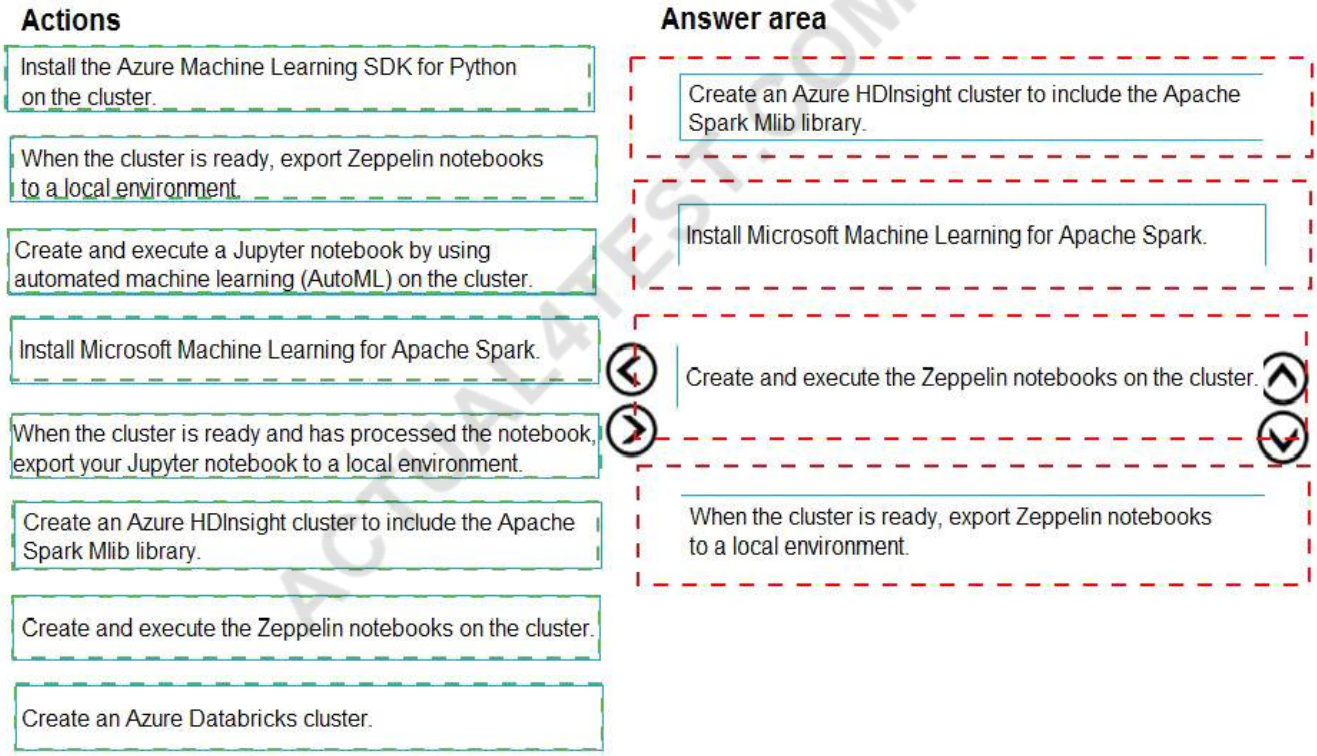


References:
https://www.edureka.co/blog/classification-algorithms/

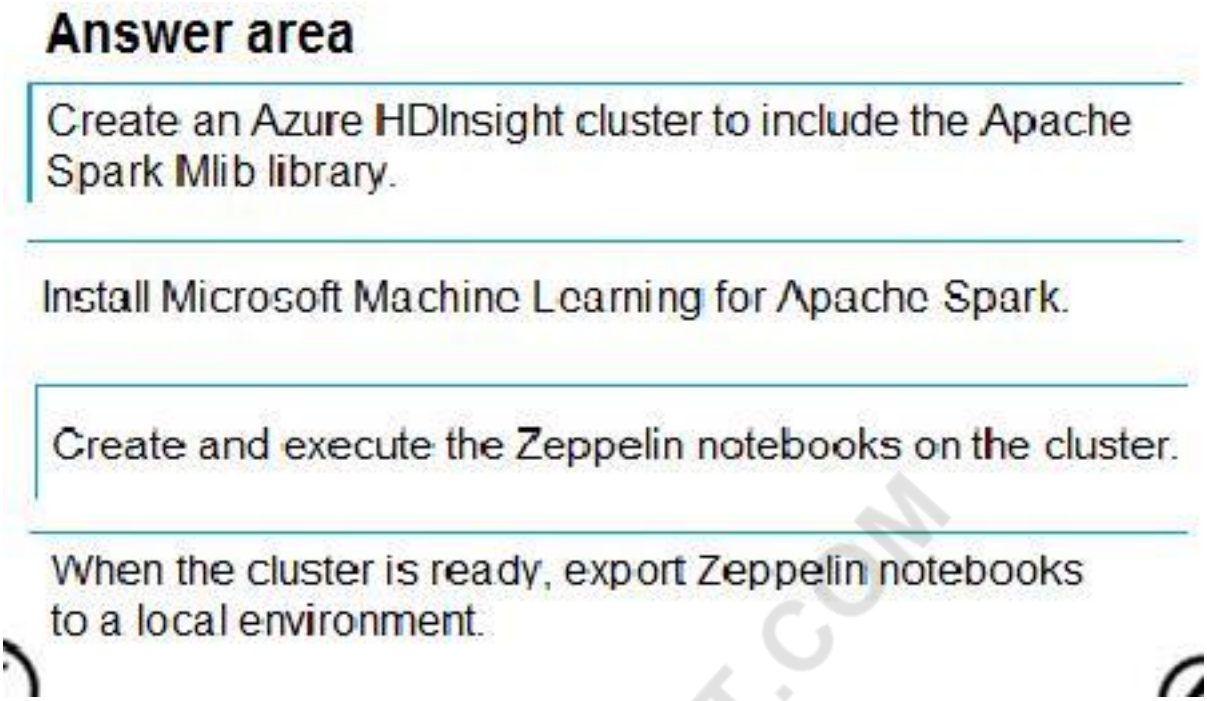
NO.28 You are building an intelligent solution using machine learning models.
The environment must support the following requirements:
* Data scientists must build notebooks in a cloud environment
* Data scientists must use automatic feature engineering and model building in machine learning pipelines.
* 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

NO.29 You are conducting feature engineering to prepuce data for further analysis.
The data includes seasonal patterns on inventory requirements.
You need to select the appropriate method to conduct feature engineering on the data.
Which method should you use?
A. Exponential Smoothing (ETS) function.
B. One Class Support Vector Machine module
C. Time Series Anomaly Detection module
D. Finite Impulse Response (FIR) Filter module.
Answer: D

NO.30 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 using Azure Machine learning Studio to perform feature engineering on a dataset. You need

to normalize values to produce a feature column grouped into bins.
Solution: Apply an Entropy Minimum Description Length (MDI) binning mode.
Does the solution meet the goal?

A. Yes

B. No

Answer: A

Explanation
Entropy MDL binning mode: This method requires that you select the column you want to predict and the column or columns that you want to group into bins. It then makes a pass over the data and attempts to determine the number of bins that minimizes the entropy. In other words, it chooses a number of bins that allows the data column to best predict the target column. It then returns the bin number associated with each row of your data in a column named <colname>quantized.
References:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/group-data-into-bins

NO.31 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 creating a model to predict the price of a student's artwork depending on the following variables: the student's length of education, degree type, and art form.
You start by creating a linear regression model.
You need to evaluate the linear regression model.
Solution: Use the following metrics: Accuracy, Precision, Recall, F1 score and AUC.
Does the solution meet the goal?

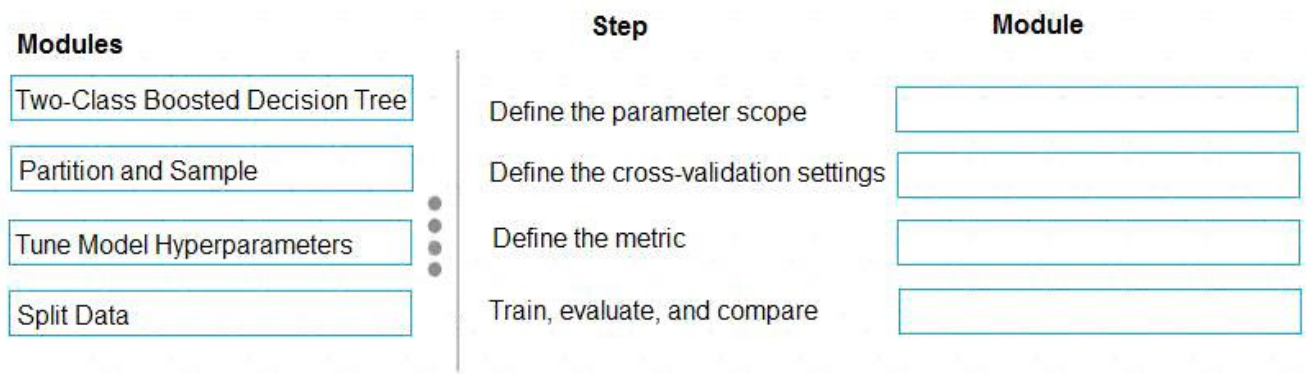
A. Yes

B. No

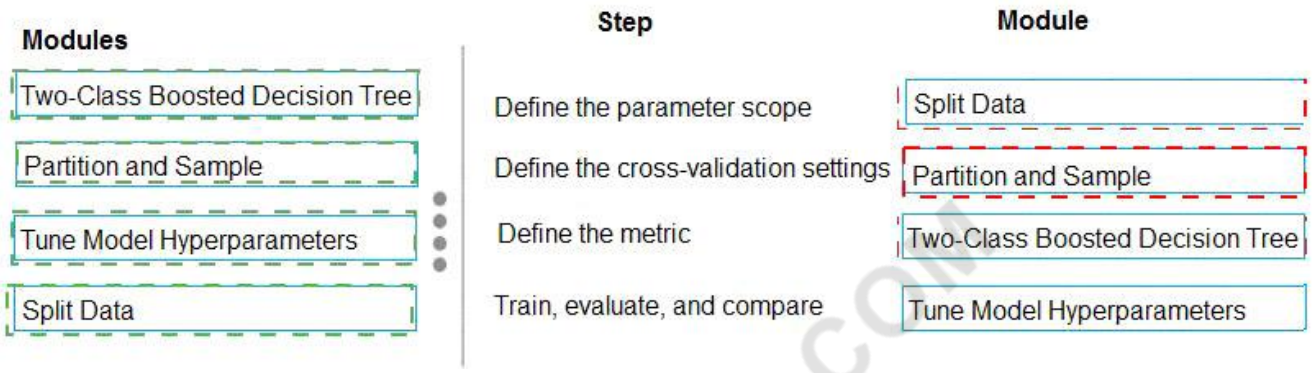
Answer: B

Explanation
Those are metrics for evaluating classification models, instead use: Mean Absolute Error, Root Mean Absolute Error, Relative Absolute Error, Relative Squared Error, and the Coefficient of Determination.
References:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/evaluate-model

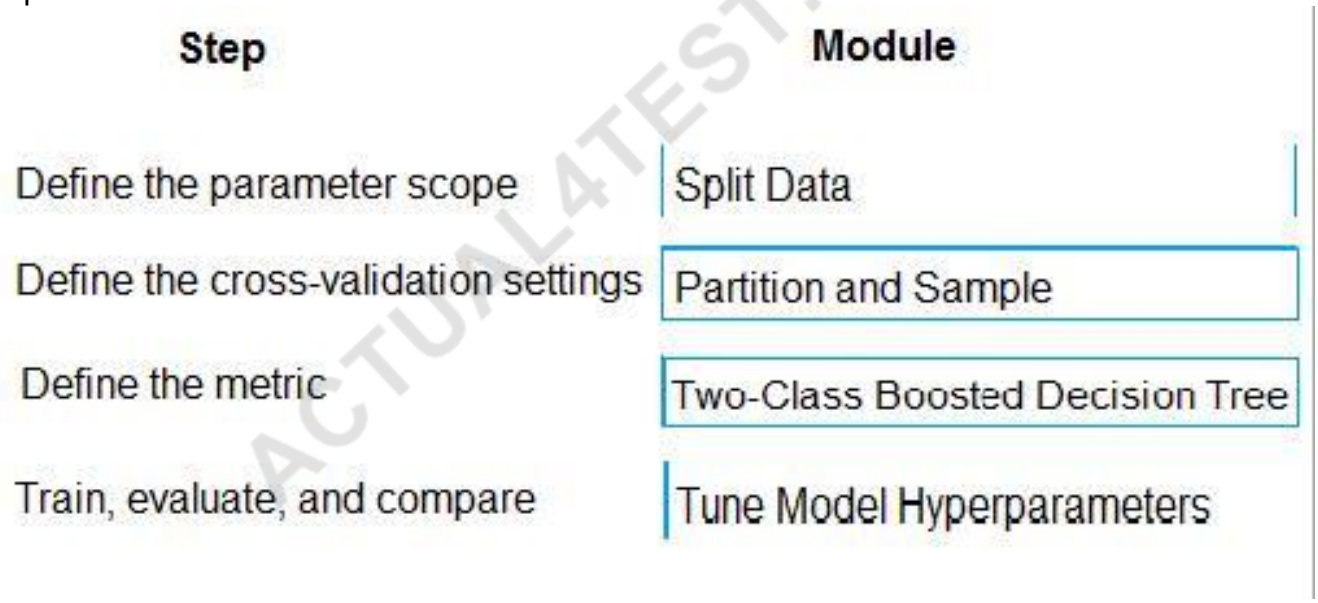
NO.32 You have a model with a large difference between the training and validation error values.
You must create a new model and perform cross-validation.
You need to identify a parameter set for the new model using Azure Machine Learning Studio.
Which module you should use for each step? To answer, 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:



Explanation

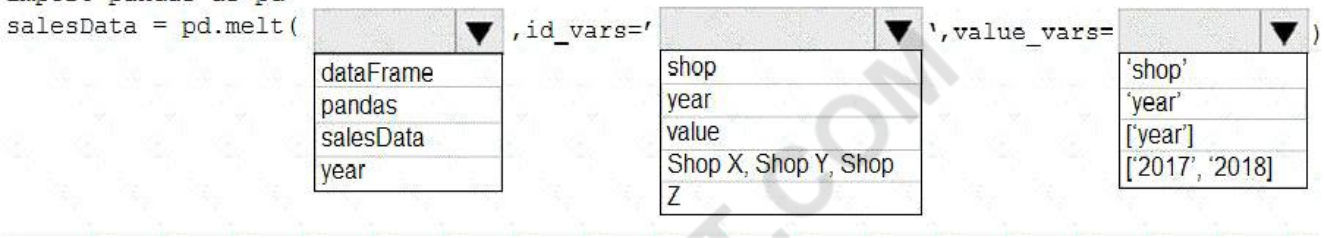


Box 1: Split data
Box 2: Partition and Sample
Box 3: Two-Class Boosted Decision Tree
Box 4: Tune Model Hyperparameters
Integrated train and tune: You configure a set of parameters to use, and then let the module iterate over multiple combinations, measuring accuracy until it finds a "best" model. With 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

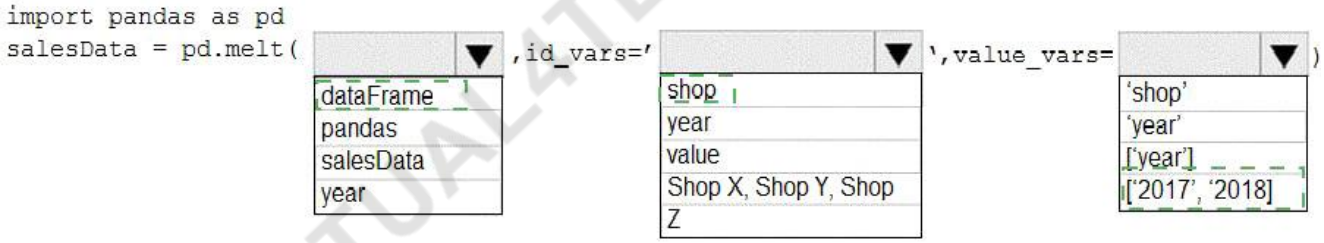
NO.33 You have a Python data frame named salesData in the following format:
The data frame must be unpivoted to a long data format as follows:
You need to use the pandas.melt() function in Python to perform the transformation.
How should you complete the code segment? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.

Answer Area

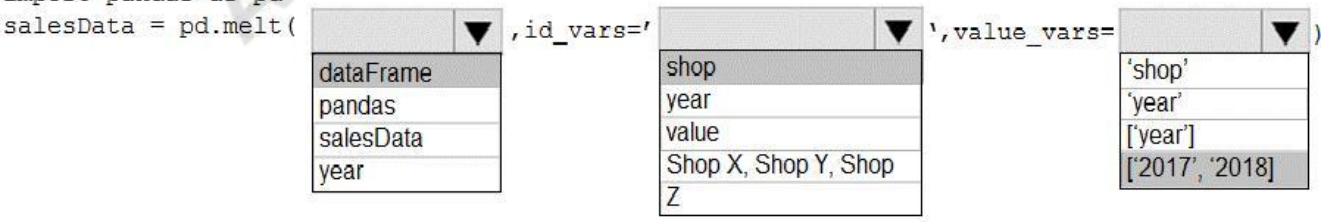


Answer:

Answer Area



Explanation



Box 1: dataFrame
Syntax: pandas.melt(frame, id_vars=None, value_vars=None, var_name=None, value_name='value', col_level=None)[source] Where frame is a DataFrame Box 2: shop Paramter id_vars id_vars : tuple, list, or ndarray, optional Column(s) to use as identifier variables.
Box 3: ['2017','2018']
value_vars : tuple, list, or ndarray, optional
Column(s) to unpivot. If not specified, uses all columns that are not set as id_vars.
Example:
df = pd.DataFrame({'A': {0: 'a', 1: 'b', 2: 'c'},
'B': {0: 1, 1: 3, 2: 5},
'C': {0: 2, 1: 4, 2: 6}})
pd.melt(df, id_vars=['A'], value_vars=['B', 'C'])

A variable value
0 a B 1
1 b B 3
2 c B 5
3 a C 2
4 b C 4
5 c C 6
References:
https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.melt.html

NO.34 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:
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.

```
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:

```
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

```
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)

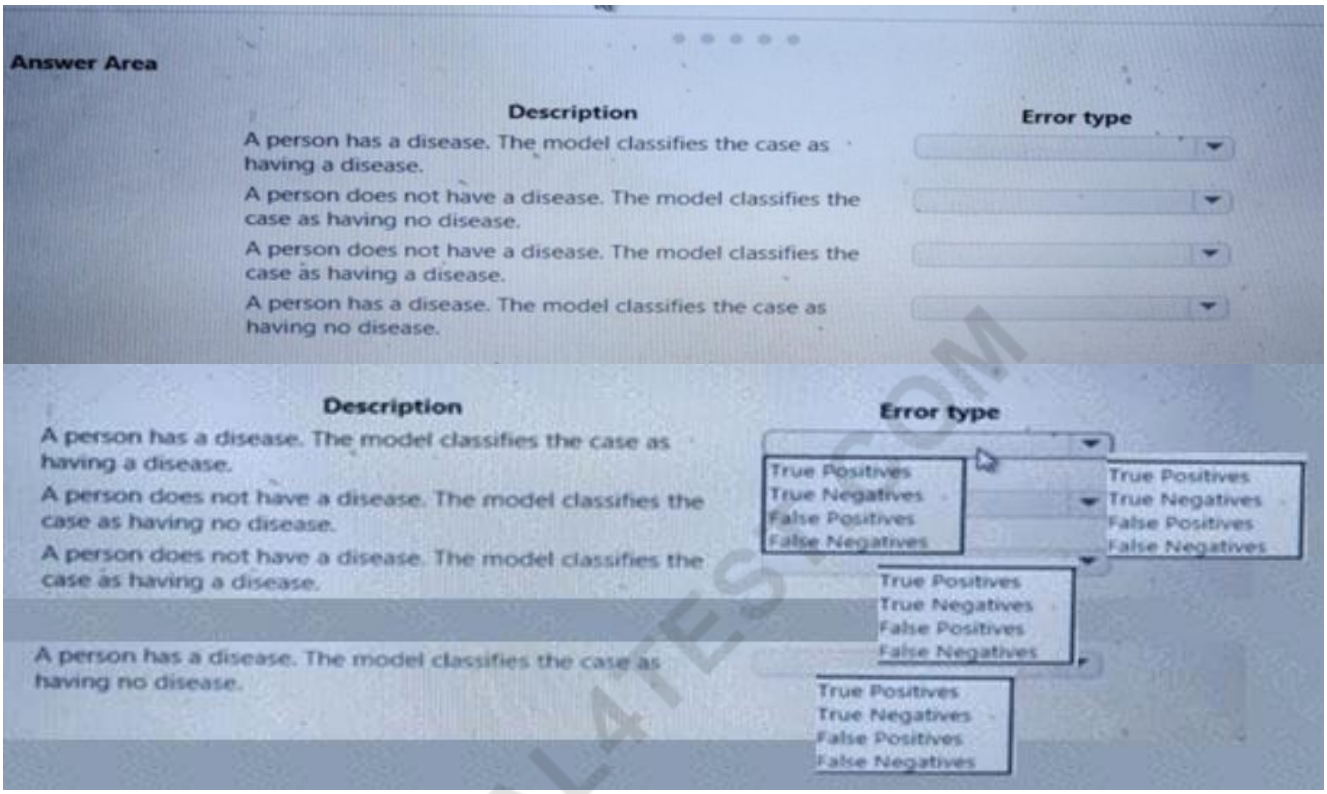
Box 1: PCA(n_components = 10)
Need to reduce the dimensionality of the feature set to 10 features in both training and testing sets.

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

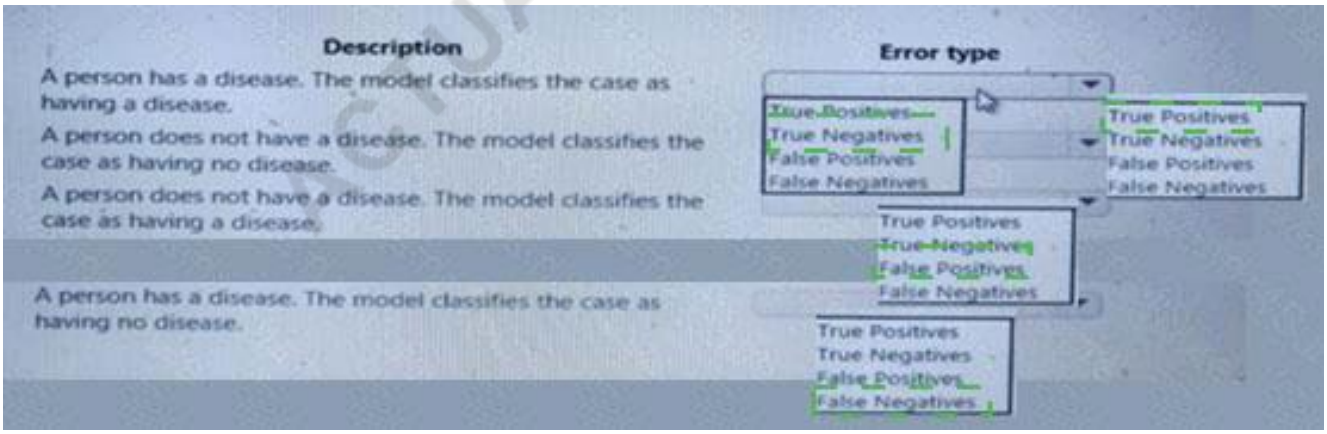
NO.35 You are building recurrent neural network to perform a binary classification.
The training loss, validation loss, training accuracy, and validation accuracy of each training epoch has been provided. You need to identify whether the classification model is over fitted.
Which of the following is correct?
A. The training loss increases while the validation loss decreases when training the model.
B. The training loss decreases while the validation loss increases when training the model.
C. The training loss stays constant and the validation loss decreases when training the model.
D. The training loss .stays constant and the validation loss stays on a constant value and close to the training loss value when training the model.
Answer: B
Explanation
An overfit model is one where performance on the train set is good and continues to improve, whereas performance on the validation set improves to a point and then begins to degrade.
References:
https://machinelearningmastery.com/diagnose-overfitting-underfitting-lstm-models/

NO.36 You are creating a machine learning model.
You need to identify outliers data.
Which two visualizations can you use? Each correct answer presents a complete solution.
NOTE: Each correct selection is worth one point.
A. box plot
B. scatter
C. random forest diagram
D. Venn diagram
E. ROC curve
Answer: A B

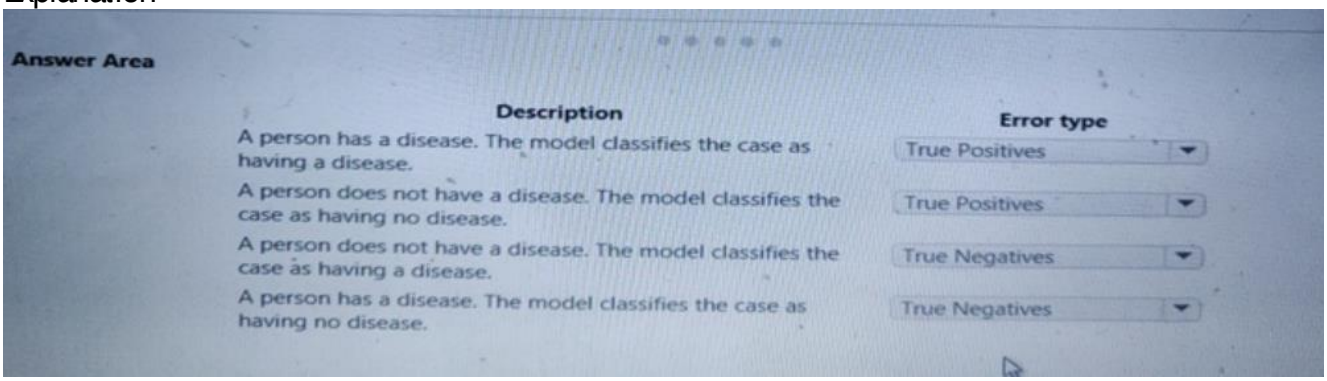
NO.37 You create a binary classification model to predict whether a person has a disease. You need to detect possible classification errors.
Which error type should you choose for each description? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.



Answer:

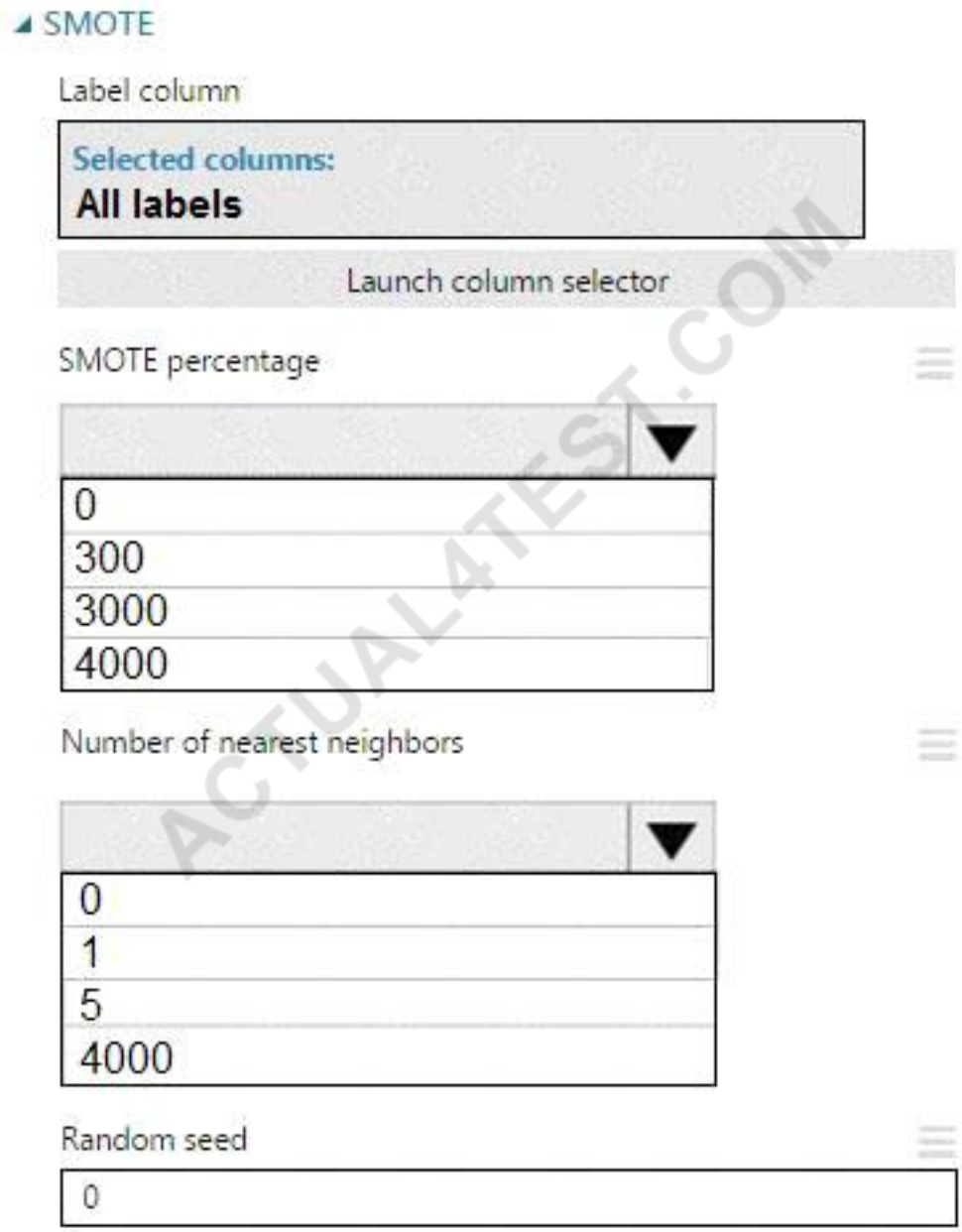


Explanation

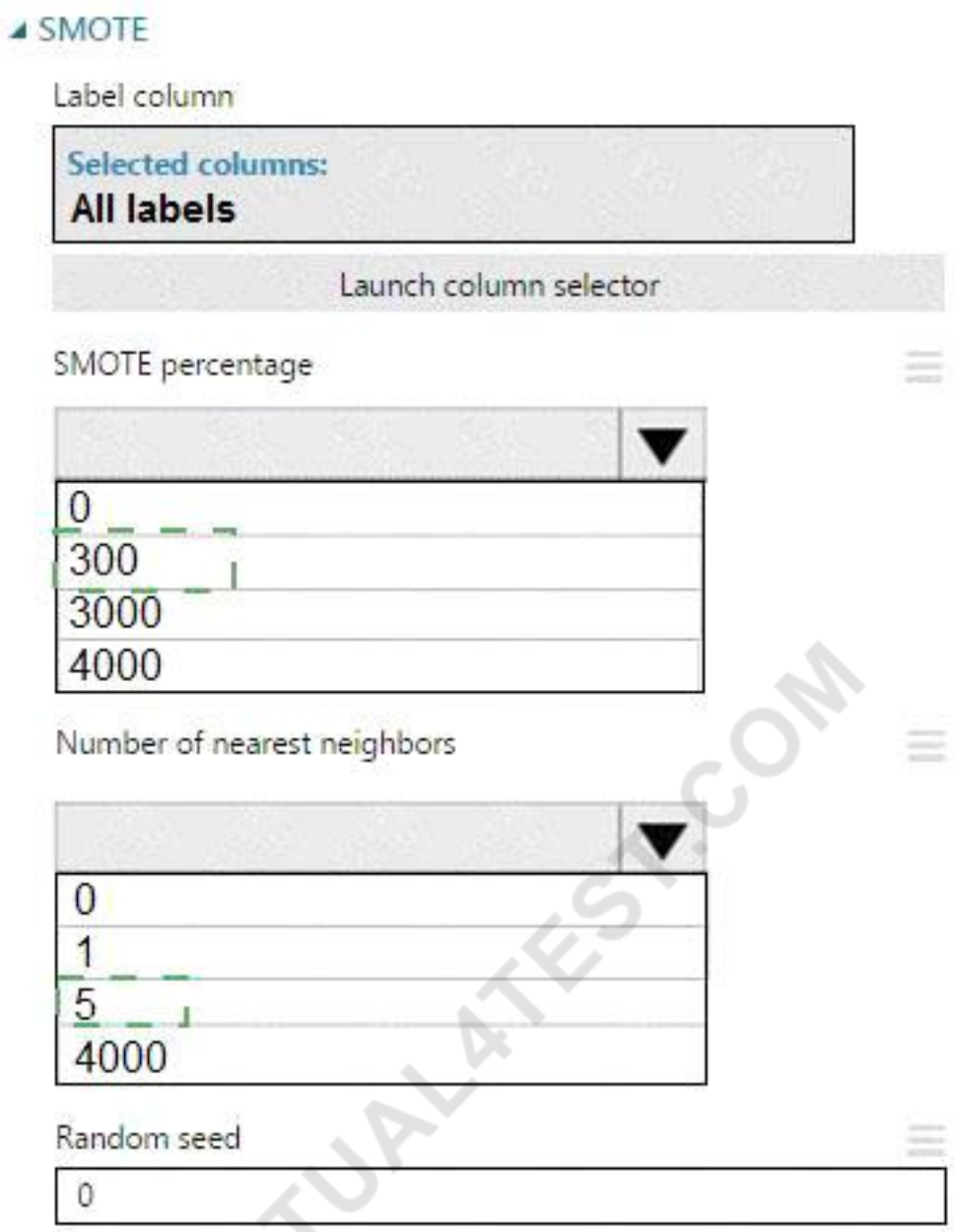


NO.38 You create an experiment in Azure Machine Learning Studio. You add a training dataset that contains 10,000 rows. The first 9,000 rows represent class 0 (90 percent).
The remaining 1,000 rows represent class 1 (10 percent).
The training set is imbalances between two classes. You must increase the number of training

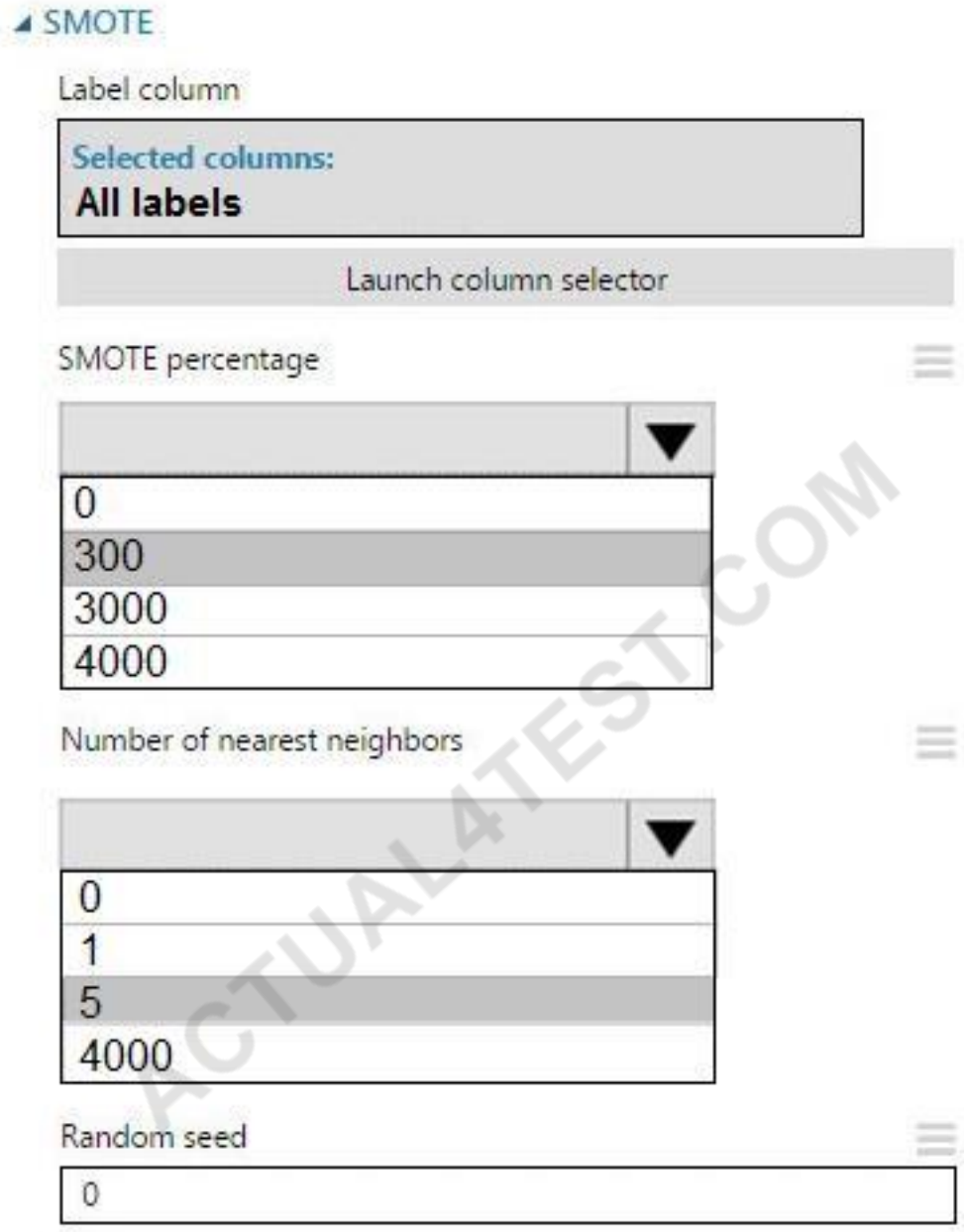
examples for class 1 to 4,000 by using 5 data rows. You add the Synthetic Minority Oversampling Technique (SMOTE) module to the experiment.
You need to configure the module.
Which values should you use? To answer, select the appropriate options in the dialog box in the answer area.
NOTE: Each correct selection is worth one point.



Answer:

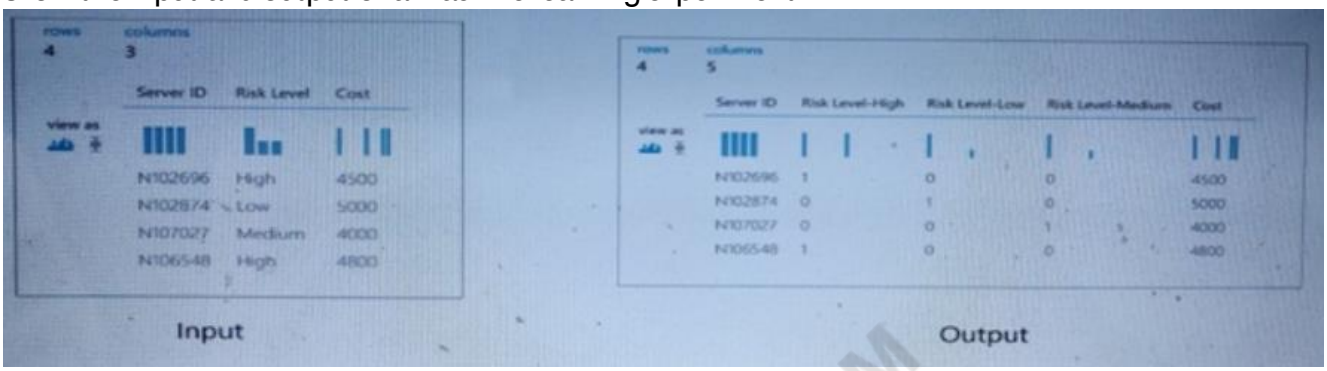


Explanation

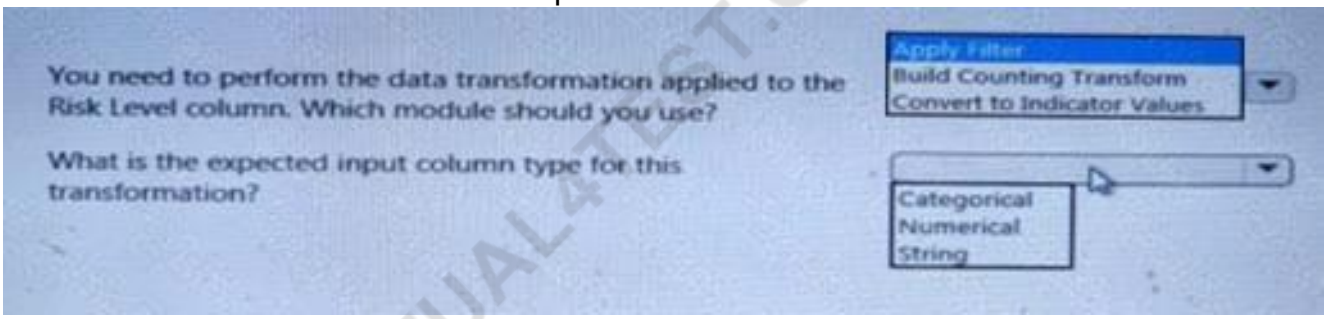


Box 1: 300
You type 300 (%), the module triples the percentage of minority cases (3000) compared to the original dataset (1000).
Box 2: 5
We should use 5 data rows.
Use the Number of nearest neighbors option to determine the size of the feature space that the SMOTE algorithm uses when in building new cases. A nearest neighbor is a row of data (a case) that is very similar to some target case. The distance between any two cases is measured by combining the weighted vectors of all features.
By increasing the number of nearest neighbors, you get features from more cases.
By keeping the number of nearest neighbors low, you use features that are more like those in the original sample.
References:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/smote

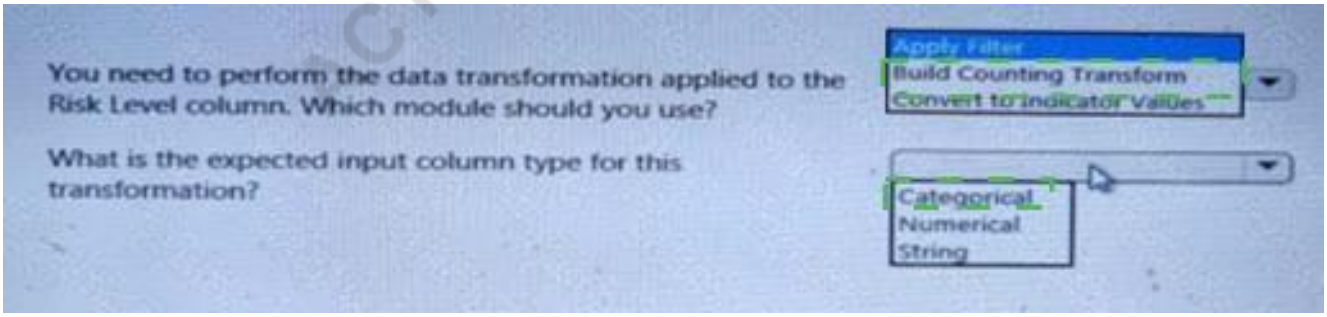
NO.39 You are developing a machine learning, experiment by using Azure. The following images show the input and output of a machine learning experiment:



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:



NO.40 You are developing a hands-on workshop to introduce Docker for Windows to attendees.
You need to ensure that workshop attendees can install Docker on their devices.
Which two prerequisite components should attendees install on the devices? Each correct answer presents part of the solution.
NOTE: Each correct selection is worth one point.

A. Microsoft Hardware-Assisted Virtualization Detection Tool

B. Kitematic

C. BIOS-enabled virtualization

D. VirtualBox

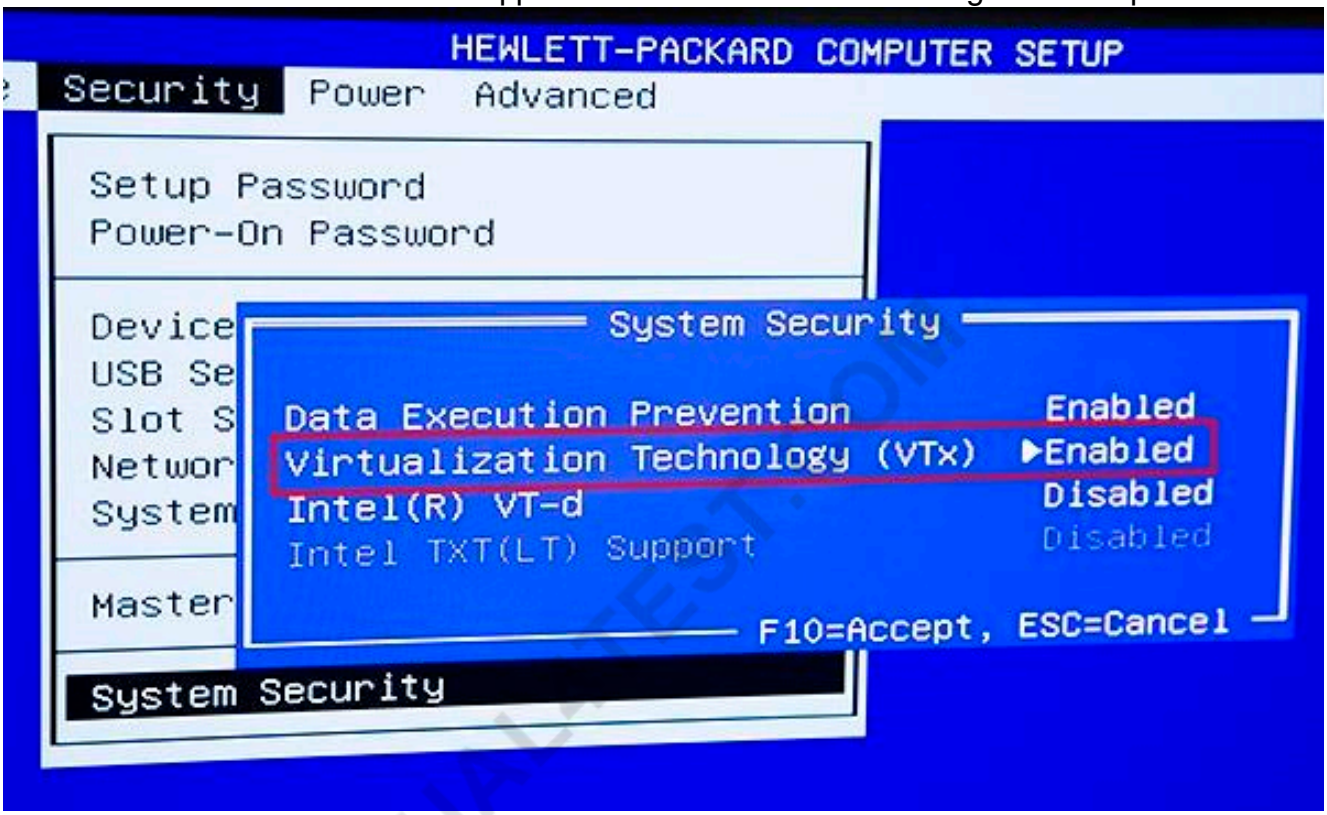
E. Windows 10 64-bit Professional

Answer: CE

Explanation

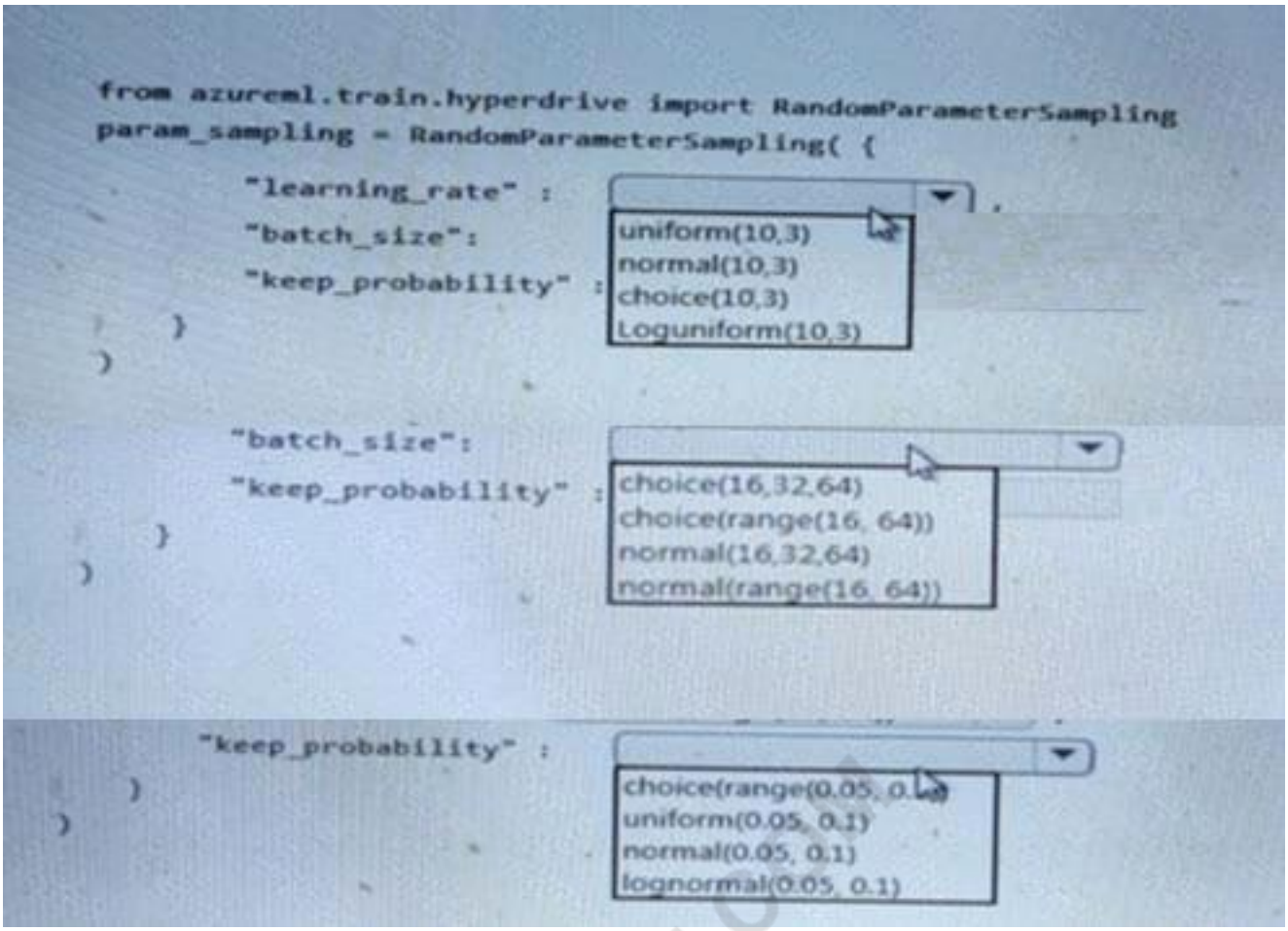
C: Make sure your Windows system supports Hardware Virtualization Technology and that virtualization is enabled.

Ensure that hardware virtualization support is turned on in the BIOS settings. For example:

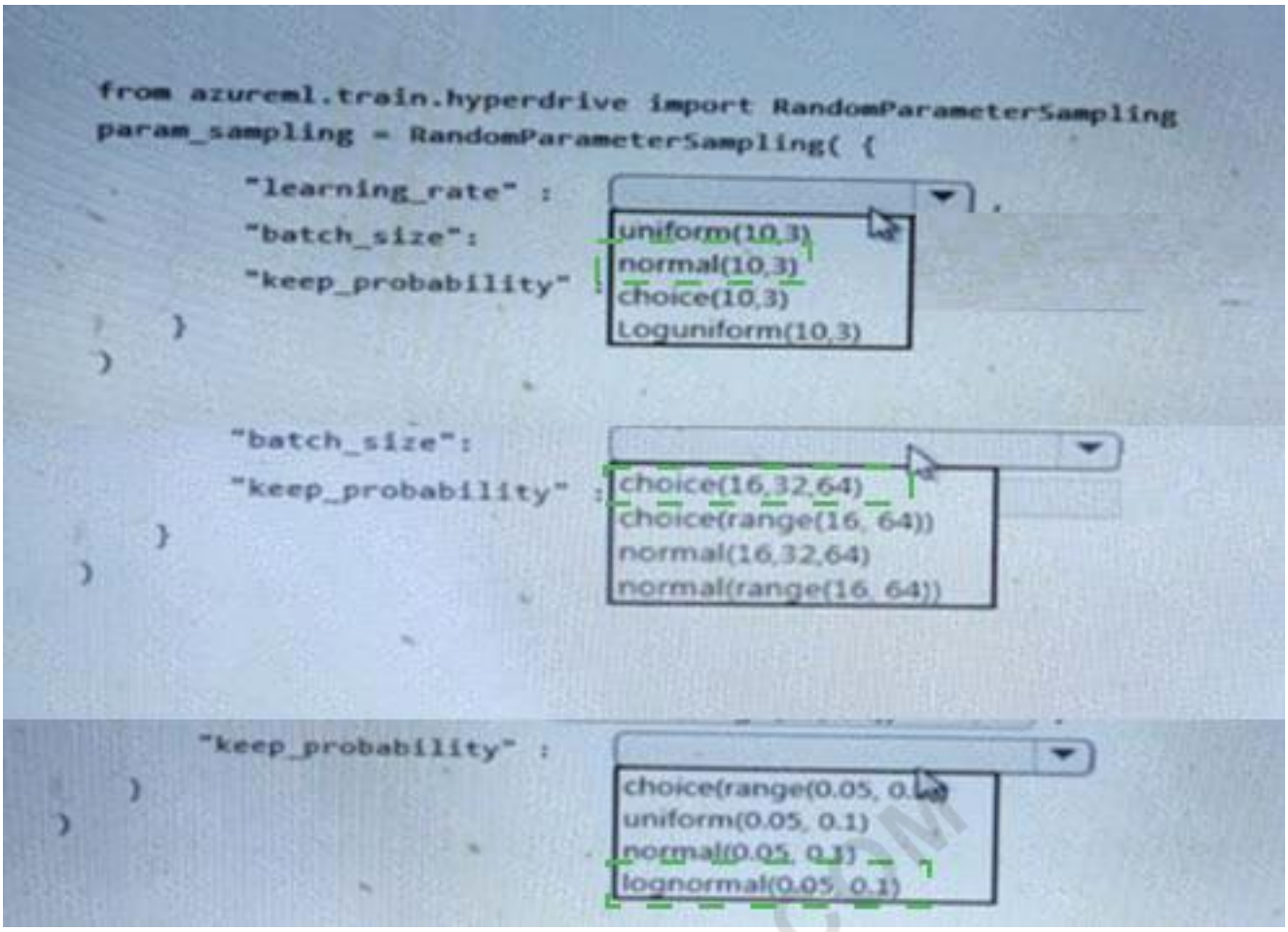


E: To run Docker, your machine must have a 64-bit operating system running Windows 7 or higher.
References:
https://docs.docker.com/toolbox/toolbox_install_windows/
https://blogs.technet.microsoft.com/canitpro/2015/09/08/step-by-step-enabling-hyper-v-for-use-on-windows-10/

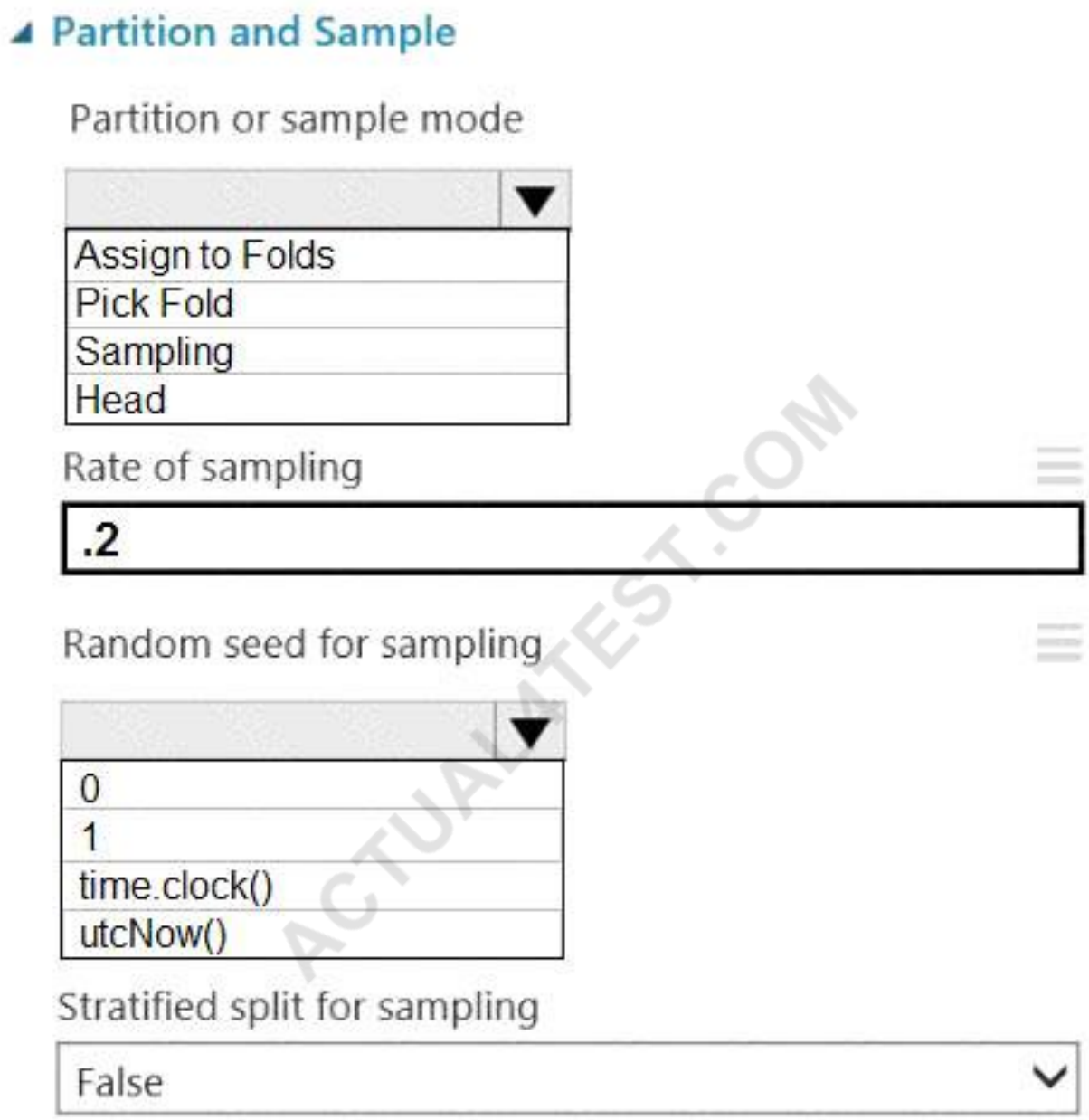
NO.41 You are using the Azure Machine Learning Service to automate hyper par a meter exploration of your neural network classification model.
You must define the hyper parameter space to automatically tune hyper parameters using random sampling according to following requirements:
* Learning rate must be selected from a normal distribution with a mean value of 10 and a standard deviation of 3.
* Batch size must be 16, 32 and 64.
* Keep probability must be a value selected from a uniform distribution between the range of 0.05 and 0.1.
You need to use the par am .sampling method of the Python API for the Azure Machine Learning Service.
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.



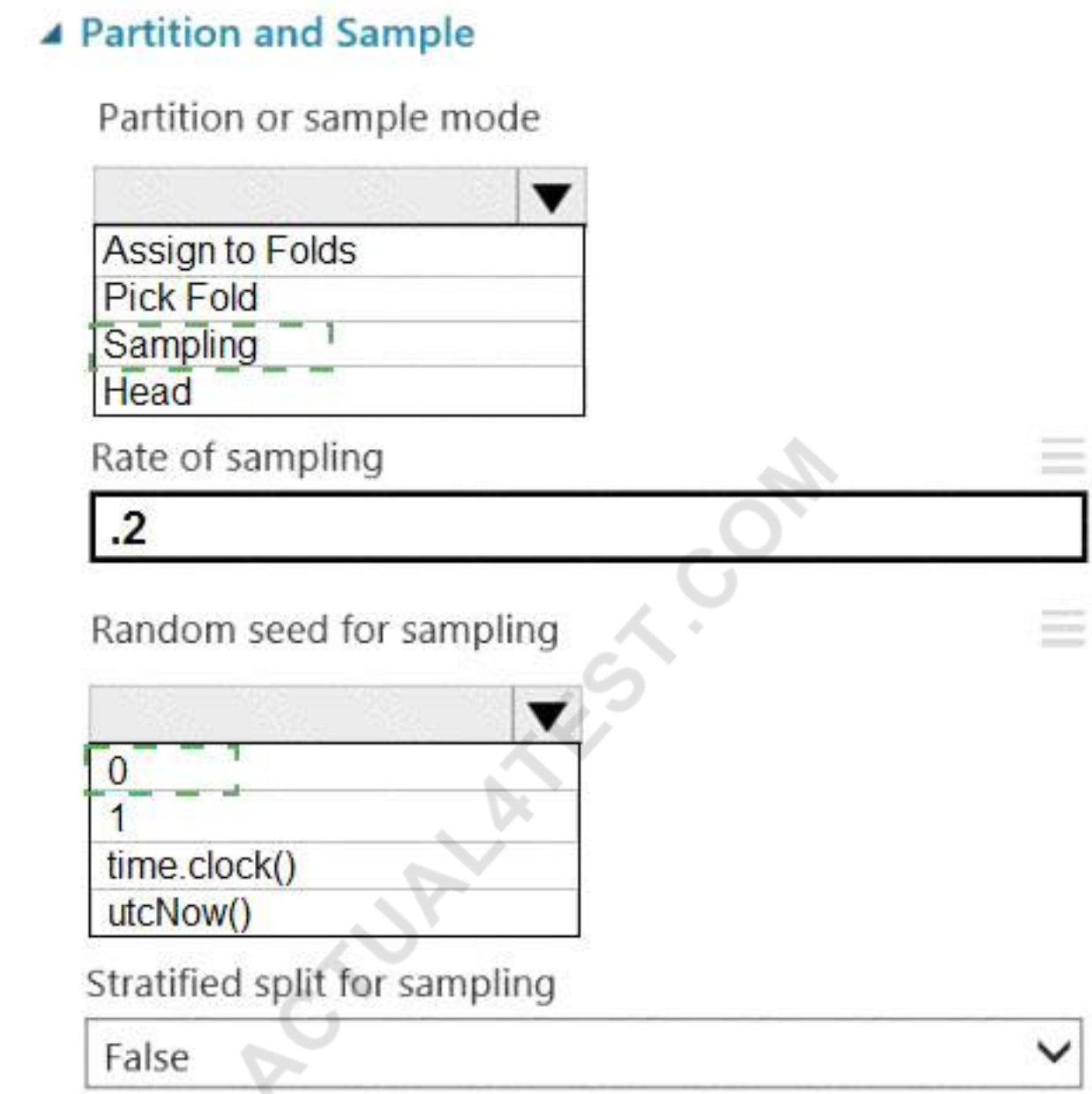
Answer:



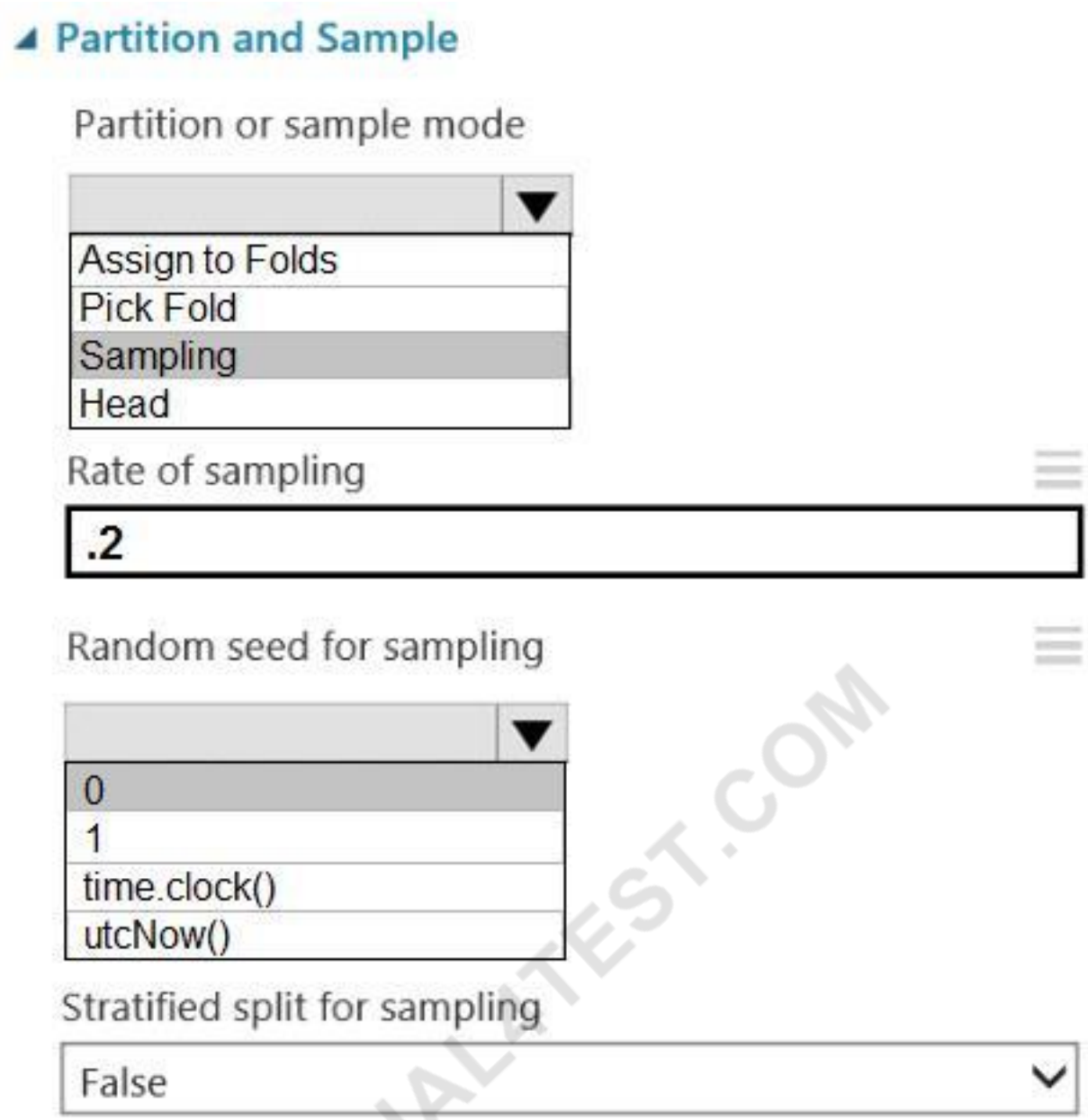
NO.42 You are retrieving data from a large datastore by using Azure Machine Learning Studio.
You must create a subset of the data for testing purposes using a random sampling seed based on the system clock.
You add the Partition and Sample module to your experiment.
You need to select the properties for the module.
Which values should you select? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.



Answer:



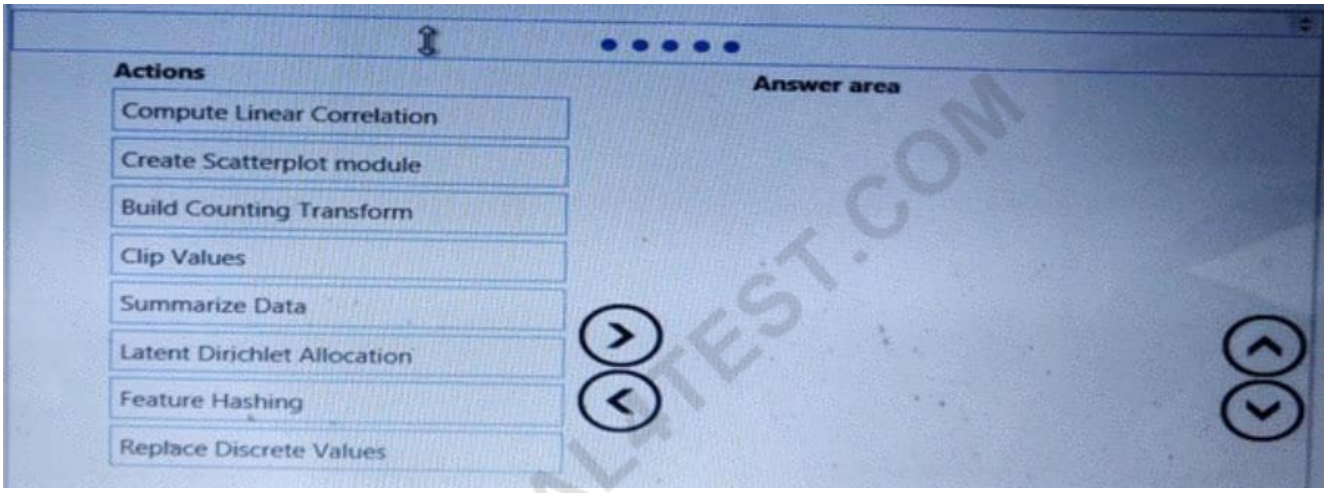
Explanation



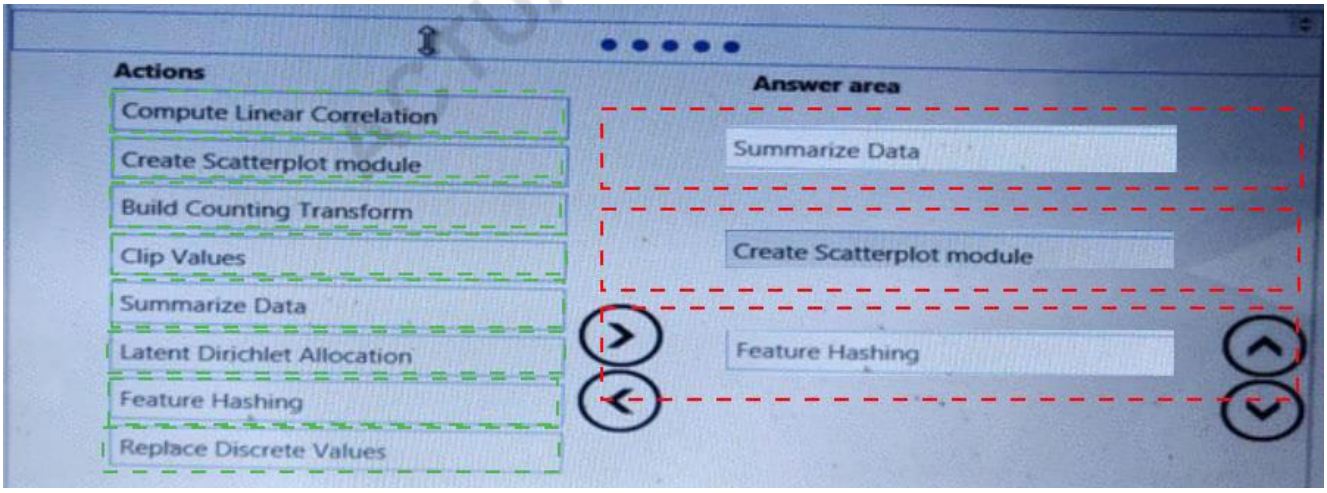
Box 1: Sampling
Create a sample of data
This option supports simple random sampling or stratified random sampling. This is useful if you want to create a smaller representative sample dataset for testing.
1. Add the Partition and Sample module to your experiment in Studio, and connect the dataset.
2. Partition or sample mode: Set this to Sampling.
3. Rate of sampling. See box 2 below.
Box 2: 0
3. Rate of sampling. Random seed for sampling: Optionally, type an integer to use as a seed value. This option is important if you want the rows to be divided the same way every time. The default value is 0, meaning that a starting seed is generated based on the system clock. This can lead to slightly different results each time you run the experiment.
References:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/partition-and-sample

NO.43 You plan to use .a Deep learning Virtual Machine (DLVM) to train deep learning models using Compute Unified Device Architecture (CUDA) computations.
You need to configure the IXVM to support CUOA
What should you implement?
A. Intel Software Guard Extensions (Intel SGX) technology
B. Solid State Drives (SSD)
C. Graphic Processing Unit (GPU)
D. Computer Processing Unit (CPU) speed increase by using overcloking
E. High Random Access Memory (RAM) configuration
Answer: B

NO.44 You need to visually identify whether outliers exist in the Age column and quantify the outliers before the outliers are removed.
Which three Azure Machine Learning Studio modules should you use in sequence? To answer, move the appropriate modules from the list of modules to the answer area and arrange them in the correct order.



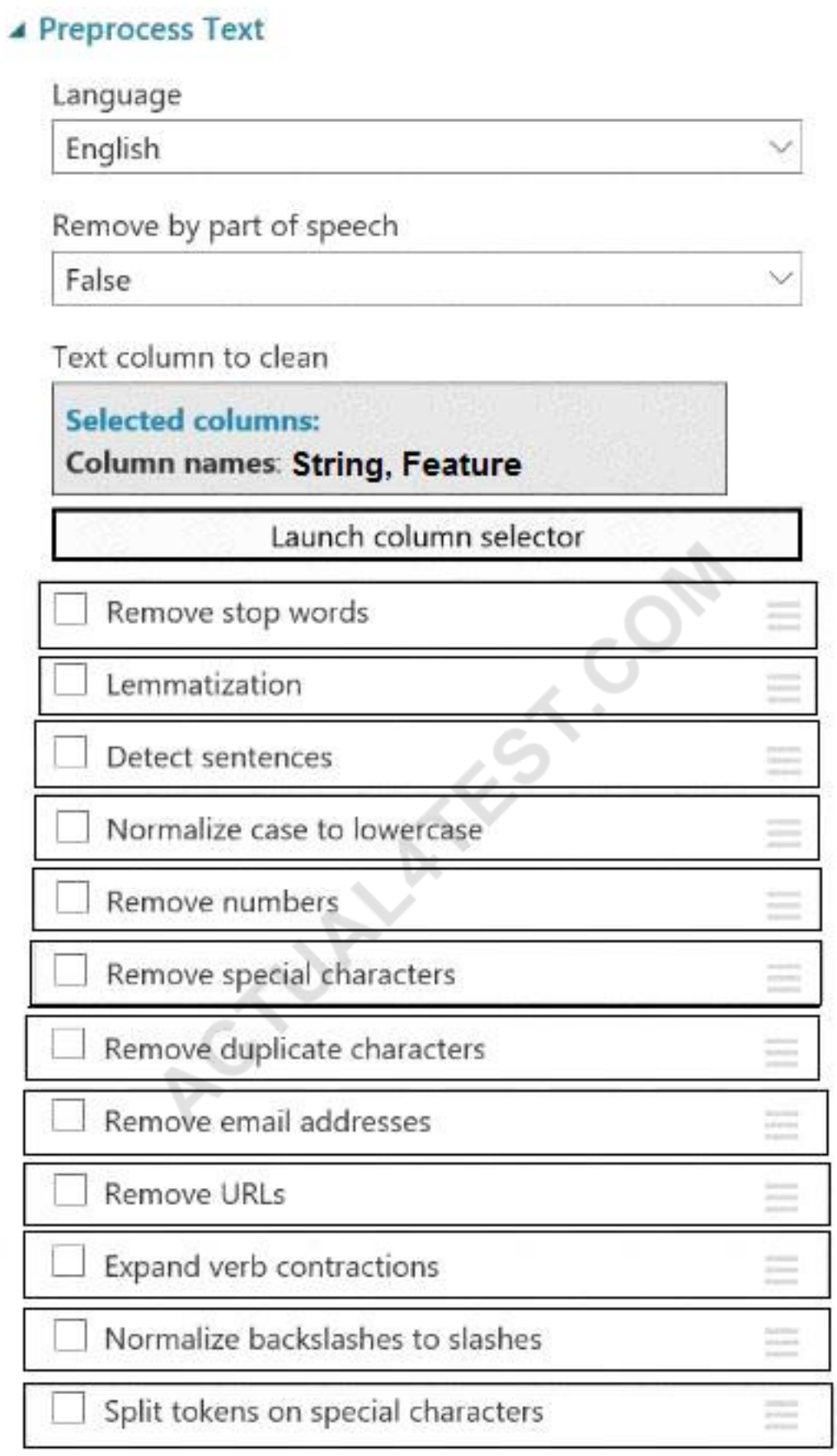
Answer:



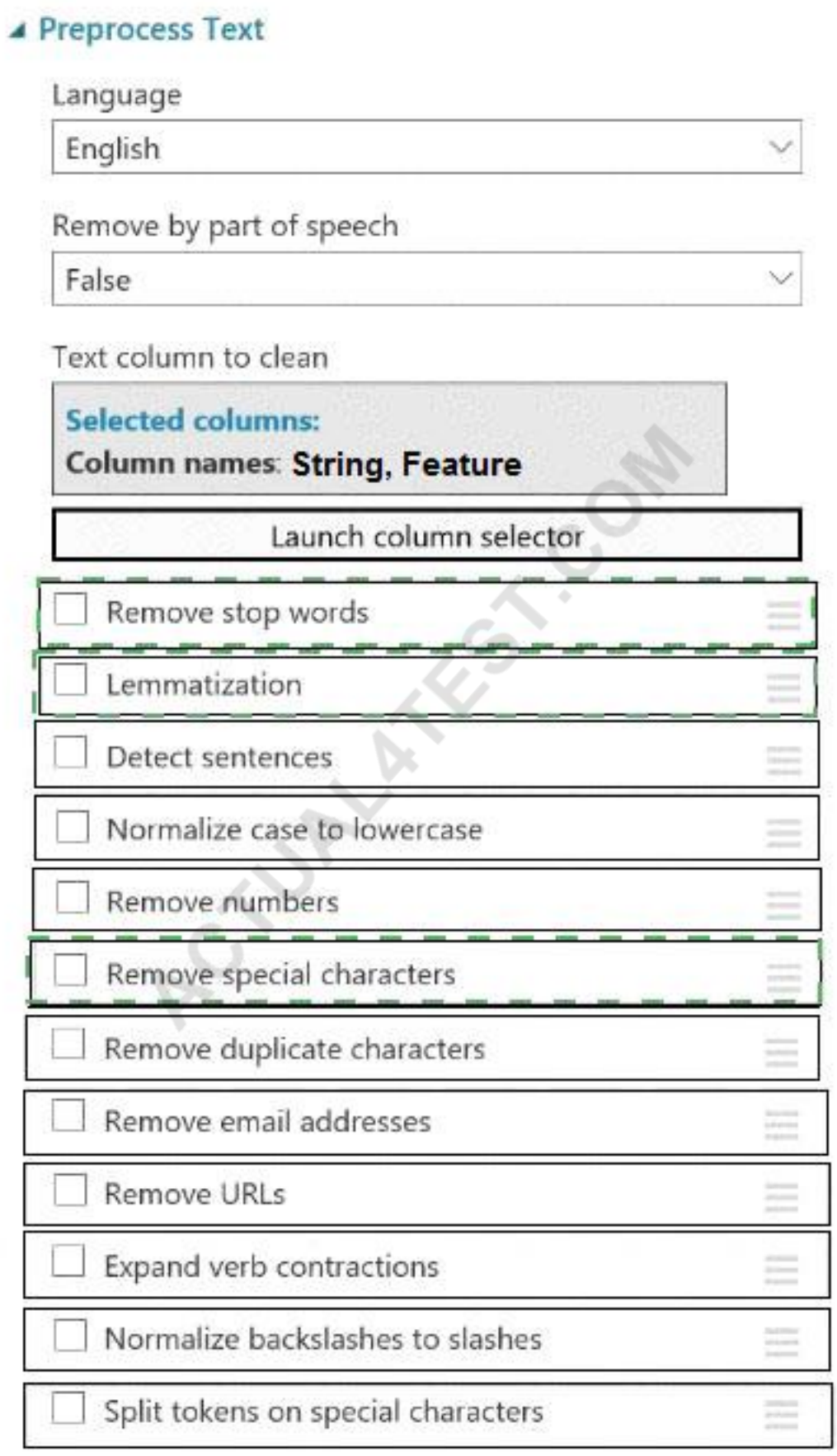
NO.45 You plan to preprocess text from CSV files. You load the Azure Machine Learning Studio default stop words list.
You need to configure the Preprocess Text module to meet the following requirements:
* Ensure that multiple related words from a single canonical form.

* Remove pipe characters from text.
* Remove words to optimize information retrieval.

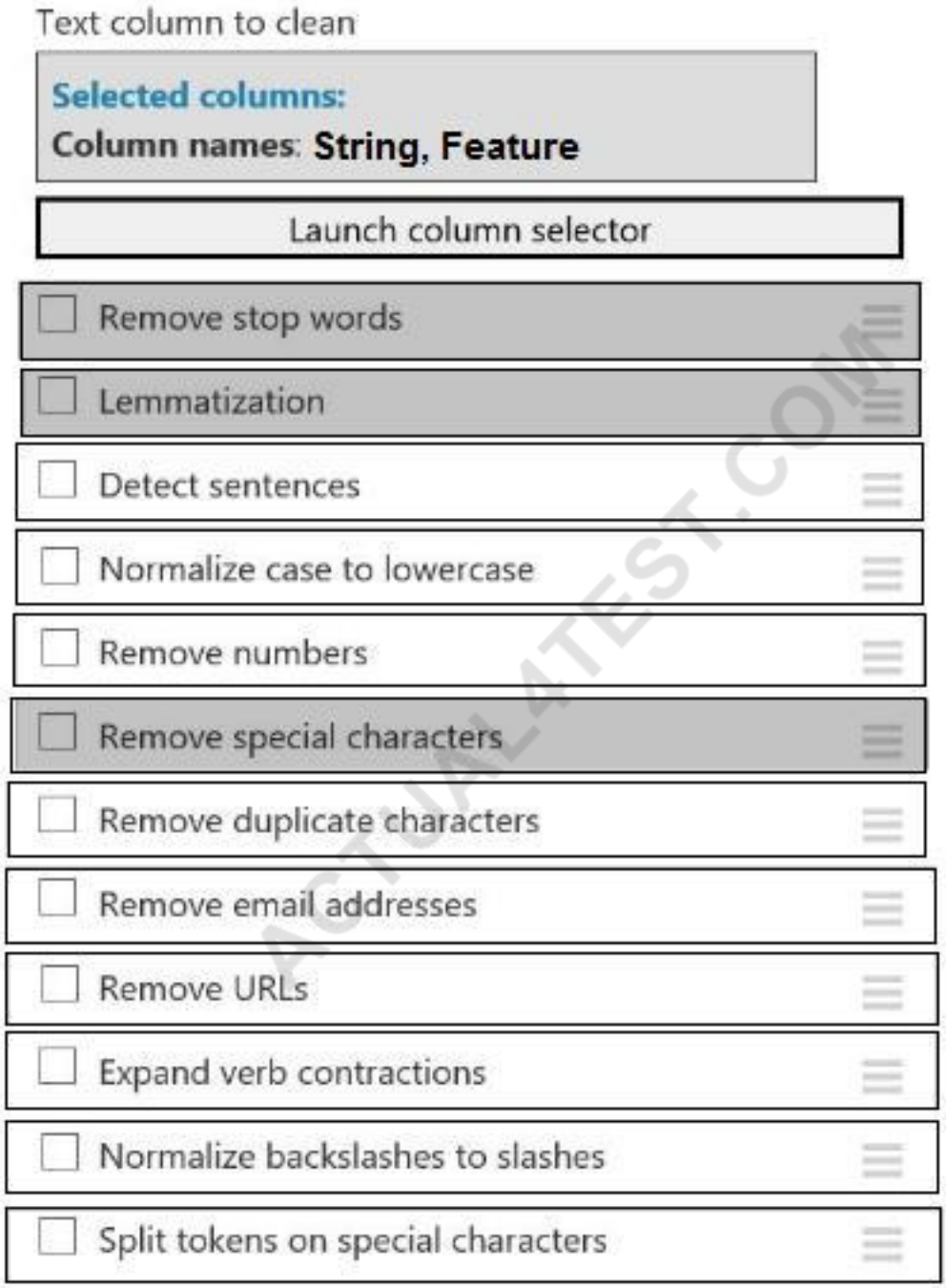
Which three options should you select? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.


Preprocess Text
Language
English
Remove by part of speech
False
Text column to clean
Selected columns:
Column names: String, Feature
Launch column selector
Remove stop words
Lemmatization
Detect sentences
Normalize case to lowercase
Remove numbers
Remove special characters
Remove duplicate characters
Remove email addresses
Remove URLs
Expand verb contractions
Normalize backslashes to slashes
Split tokens on special characters

Answer:



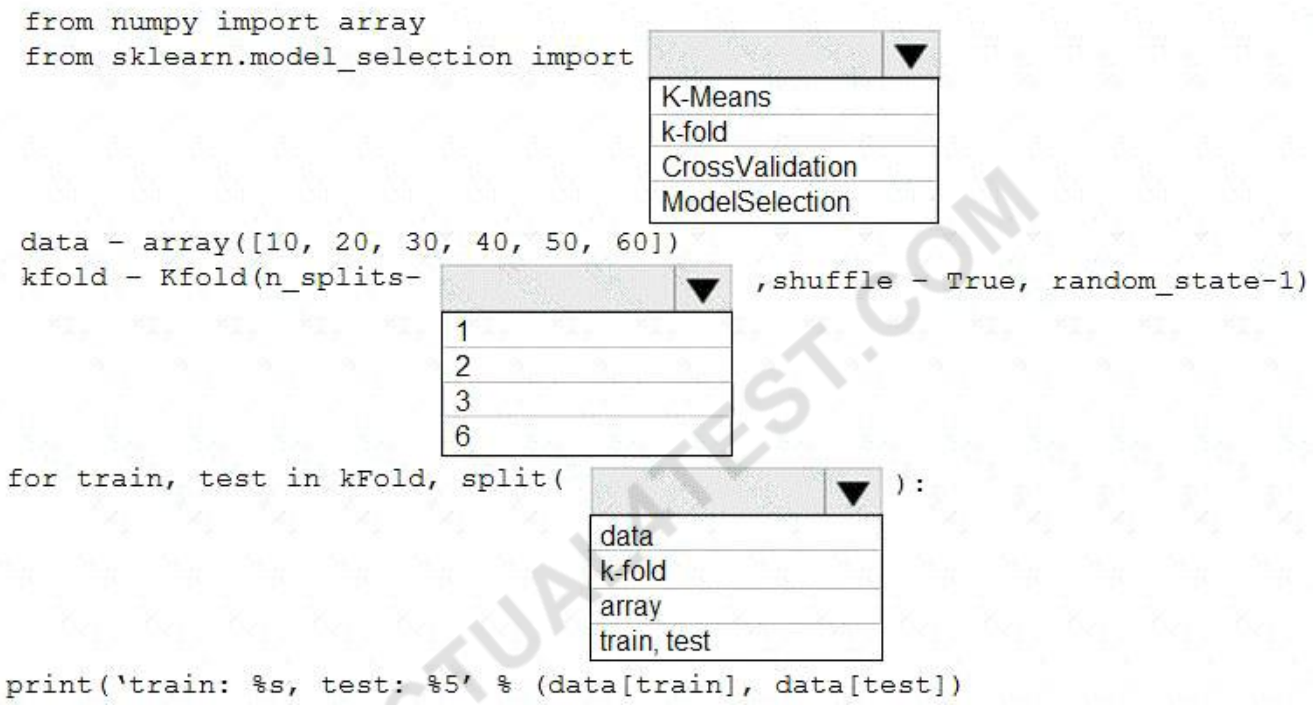
Explanation



Box 1: Remove stop words
Remove words to optimize information retrieval.
Remove stop words: Select this option if you want to apply a predefined stopword list to the text column. Stop word removal is performed before any other processes.
Box 2: Lemmatization
Ensure that multiple related words from a single canonical form.
Lemmatization converts multiple related words to a single canonical form Box 3: Remove special characters Remove special characters: Use this option to replace any non-alphanumeric special characters with the pipe | character.
References:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/preprocess-text

NO.46 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:

```
from numpy import array
from sklearn.model_selection import [ ]
```

K-Means
k-fold
CrossValidation
ModelSelection

```
data - array([10, 20, 30, 40, 50, 60])
kfold - Kfold(n_splits- [ ] ,shuffle - True, random_state-1)
```

1
2
3
6

```
for train, test in kFold, split( [ ] ):
```

data
k-fold
array
train, test

```
print('train: %s, test: %5' % (data[train], data[test])
```

Explanation

```
from numpy import array
from sklearn.model_selection import [ ]
```

K-Means
k-fold
CrossValidation
ModelSelection

```
data - array([10, 20, 30, 40, 50, 60])
kfold - Kfold(n_splits- [ ] ,shuffle - True, random_state-1)
```

1
2
3
6

```
for train, test in kFold, split( [ ] ):
```

data
k-fold
array
train, test

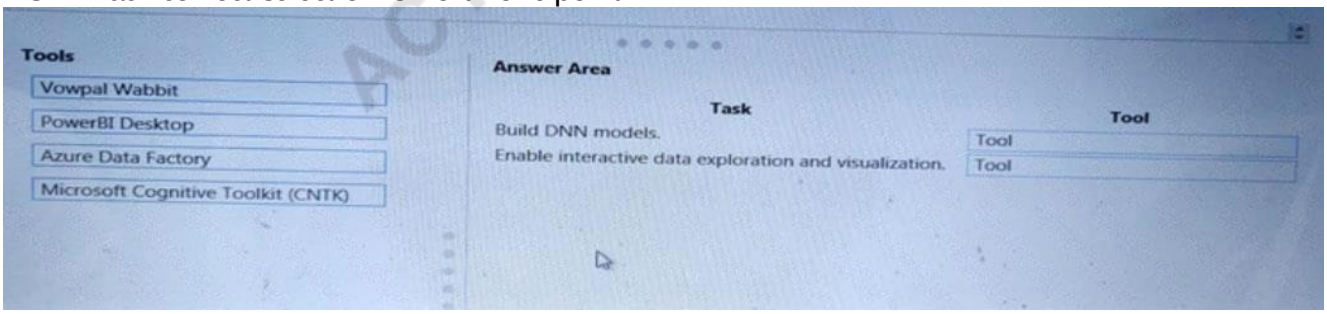
```
print('train: %s, test: %5' % (data[train], data[test])
```

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]])

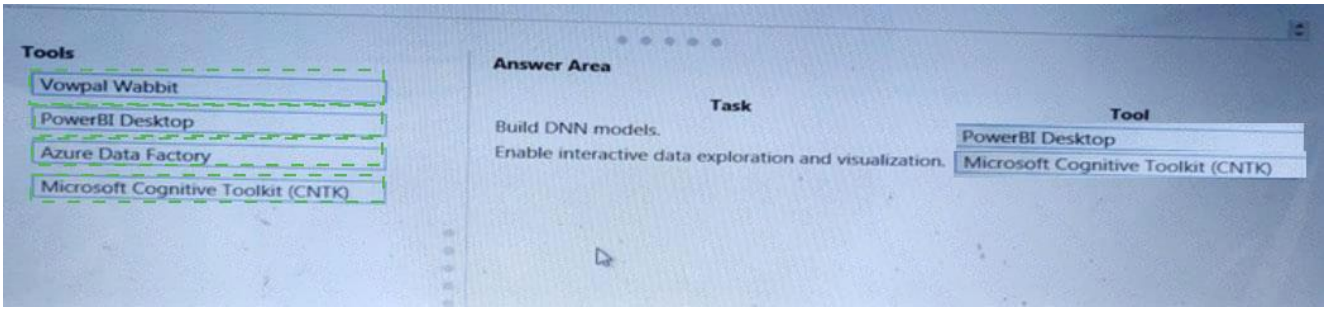
```
>>> y = np.array([1, 2, 3, 4])
>>> kf = KFold(n_splits=2)
>>> kf.get_n_splits(X)
2
>>> print(kf)
KFold(n_splits=2, random_state=None, shuffle=False)
>>> for train_index, test_index in kf.split(X):
print("TRAIN:", train_index, "TEST:", test_index)
X_train, X_test = X[train_index], X[test_index]
y_train, 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

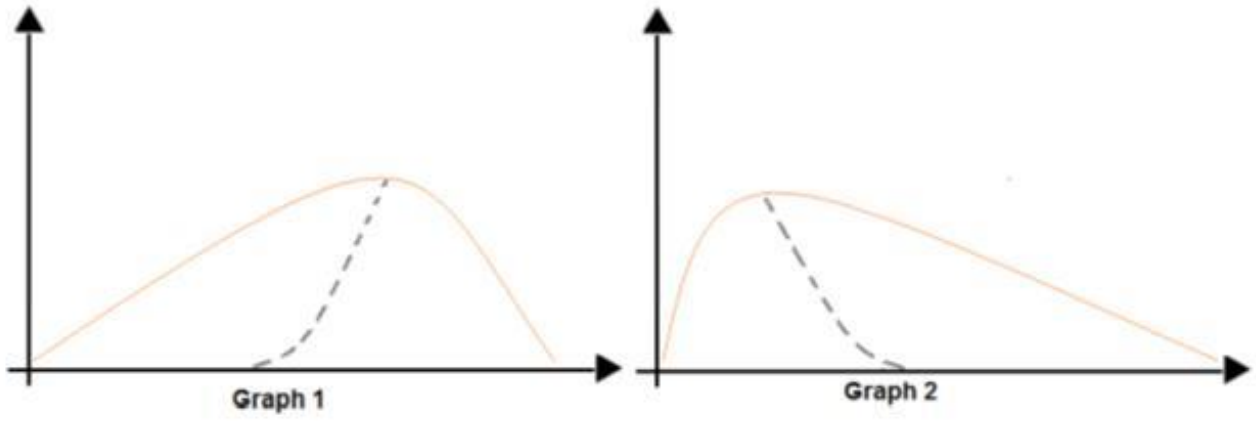
NO.47 You configure a Deep Learning Virtual Machine for Windows.
You need to recommend tools and frameworks to perform the following:
Build deep rwur.il network (DNN) models.
Perform interactive data exploration and visualization.
Which tools and frameworks should you recommend? To answer, drag the appropriate tools to the correct tasks. Each tool may be used once, 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:



NO.48 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.

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:

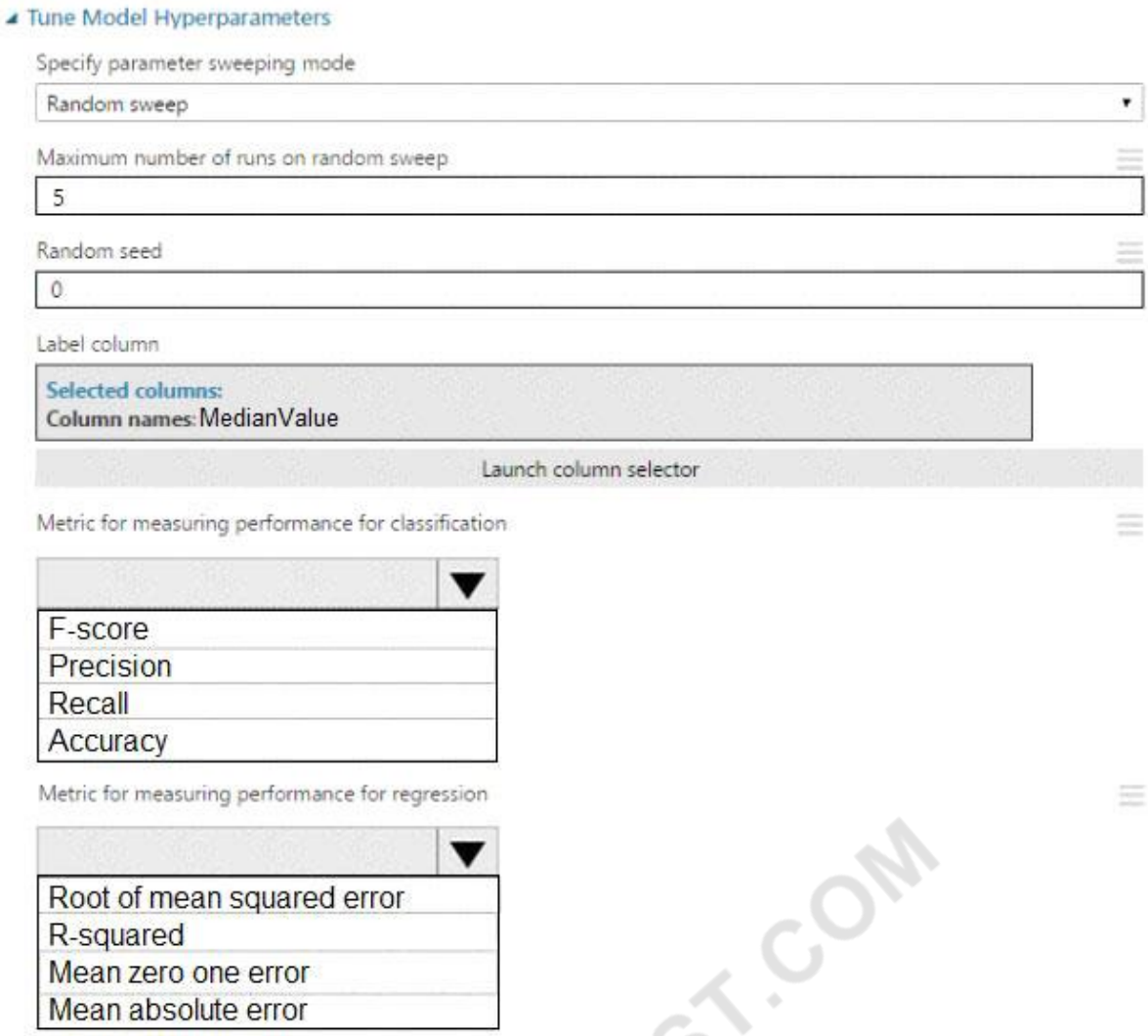
Question	Answer choice
Which type of distribution is shown for the dataset density curve of Graph 1?	Negative skew Positive skew Normal distribution Bimodal distribution
Which type of distribution is shown for the dataset density curve of Graph 2?	Negative skew Positive skew Normal distribution Bimodal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

Explanation

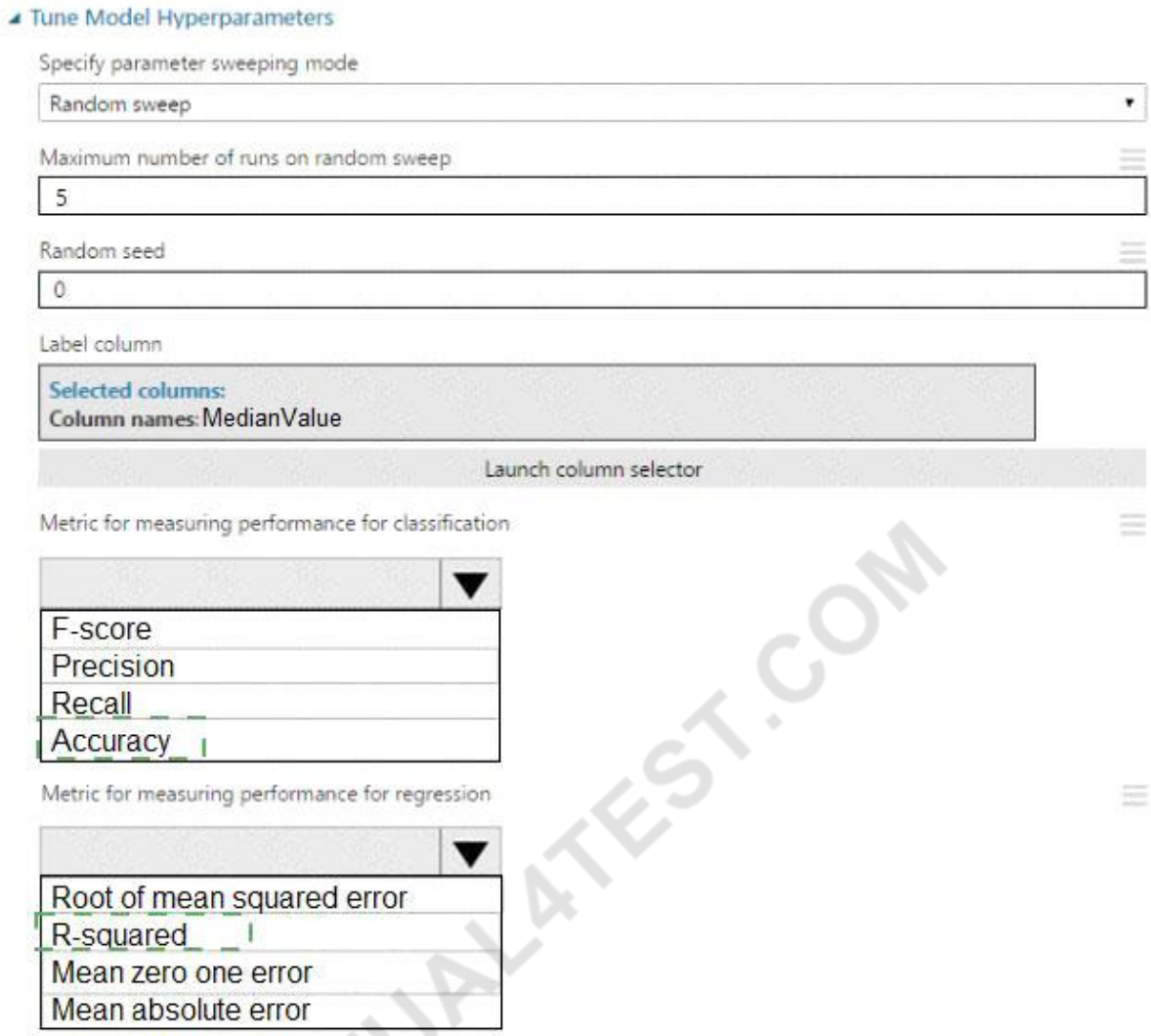
Question	Answer choice
Which type of distribution is shown for the dataset density curve of Graph 1?	Negative skew **Positive skew** Normal distribution Bimodal distribution
Which type of distribution is shown for the dataset density curve of Graph 2?	**Negative skew** Positive skew Normal distribution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

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

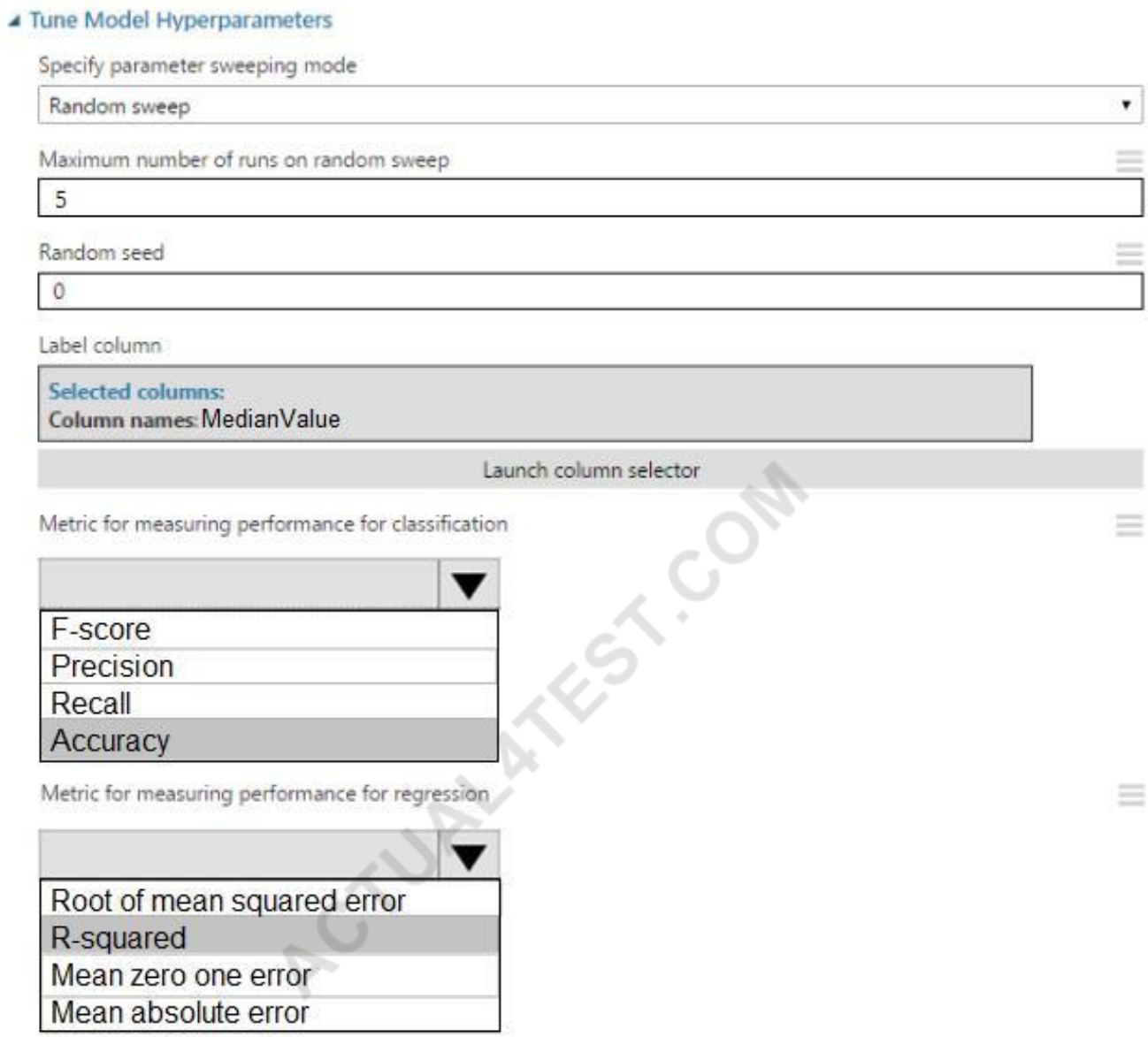
NO.49 You need to set up the Permutation Feature Importance module according to the model training requirements.
Which properties should you select? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.



Answer:



Explanation



Box 1: Accuracy
Scenario: You want to configure hyperparameters in the model learning process to speed the learning phase by using hyperparameters. In addition, this configuration should cancel the lowest performing runs at each evaluation interval, thereby directing effort and resources towards models that are more likely to be successful.
Box 2: R-Squared

NO.50 You are building a regression model tot estimating the number of calls during an event. You need to determine whether the feature values achieve the conditions to build a Poisson regression model.
Which two conditions must the feature set contain? I ach correct answer presents part of the solution. NOTE:
Each correct selection is worth one point.

A. The label data must be a negative value.

B. The label data can be positive or negative,

C. The label data must be a positive value

D. The label data must be non discrete.

E. The data must be whole numbers.

Answer: CE

Explanation

Poisson regression is intended for use in regression models that are used to predict numeric values, typically counts. Therefore, you should use this module to create your regression model only if the values you are trying to predict fit the following conditions:

* The response variable has a Poisson distribution.
* Counts cannot be negative. The method will fail outright if you attempt to use it with negative labels.
* A Poisson distribution is a discrete distribution; therefore, it is not meaningful to use this method with non-whole numbers.

References:

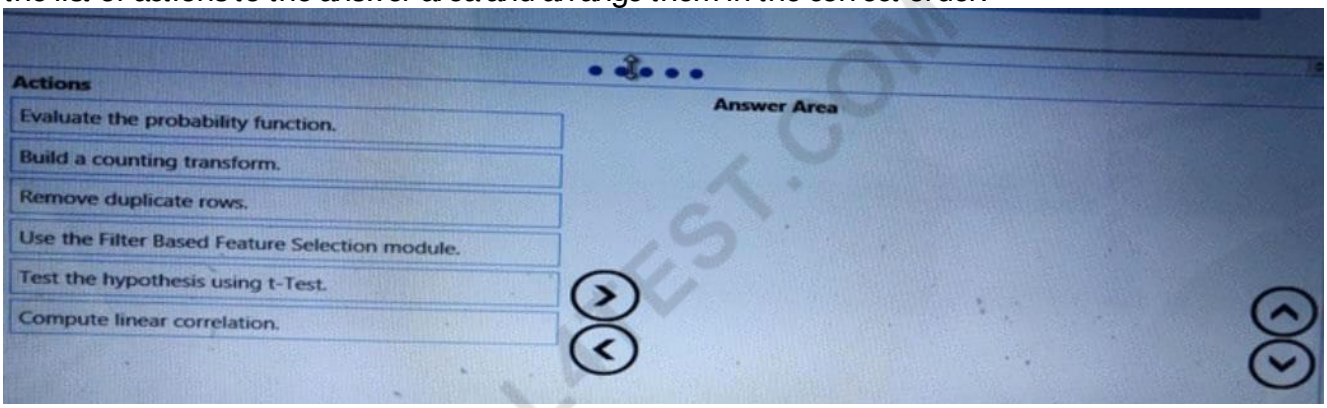
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/poisson-regression

NO.51 You are producing a multiple linear regression model in Azure Machine learning Studio.

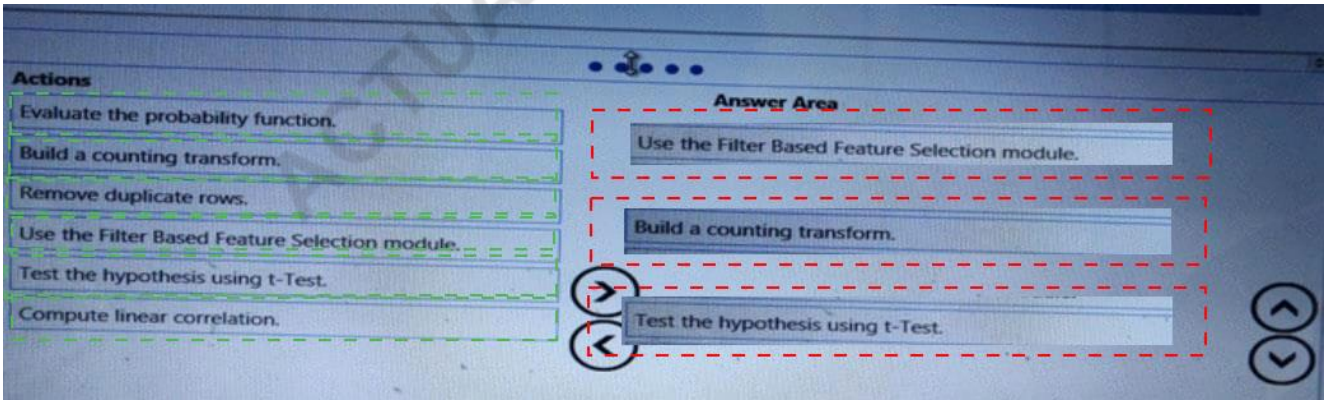
Several independent variables are highly correlated.

You need to select appropriate methods for conducting elective feature engineering on all the data.

Which three 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:



NO.52 You are performing clustering by using the K-means algorithm.

You need to define the possible termination conditions.

Which three conditions can you use? Each correct answer presents a complete solution.

NOTE: Each correct selection is worth one point.

A. A fixed number of iterations is executed.

B. The residual sum of squares (RSS) rises above a threshold.
C. The sum of distances between centroids reaches a maximum.
D. The residual sum of squares (RSS) falls below a threshold.
E. Centroids do not change between iterations.
Answer: A B C

NO.53 You are determining if two sets of data are significantly different from one another by using Azure Machine Learning Studio.
Estimated values in one set of data may be more than or less than reference values in the other set of data. You must produce a distribution that has a constant. Type I error as a function of the correlation.
You need to produce the distribution.
Which type of distribution should you produce?
A. Paired t-test with a two-tail option
B. Unpaired t-test with a two tail option
C. Paired t-test with a one-tail option
D. Unpaired t-test with a one-tail option
Answer: D

NO.54 You are creating a machine learning model. You have a dataset that contains null rows.
You need to use the Clean Missing Data module in Azure Machine Learning Studio to identify and resolve the null and missing data in the dataset.
Which parameter should you use?
A. Replace with mean
B. Remove entire column
C. Remove entire row
D. Hot Deck
Answer: B
Explanation
Remove entire row: Completely removes any row in the dataset that has one or more missing values. This is useful if the missing value can be considered randomly missing.
References:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/clean-missing-data

NO.55 You are analyzing a dataset containing historical data from a local taxi company. You arc developing a regression a regression model.
You must predict the fare of a taxi trip.
You need to select performance metrics to correctly evaluate the- regression model.
Which two metrics can you use? Each correct answer presents a complete solution.
NOTE: Each correct selection is worth one point.
A. an F1 score that is high
B. an R Squared value dose to 1
C. an R-Squared value close to 0

D. a Root Mean Square Error value that is high
E. a Root Mean Square Error value that is tow
F. an F 1 score that is low.
Answer: D F

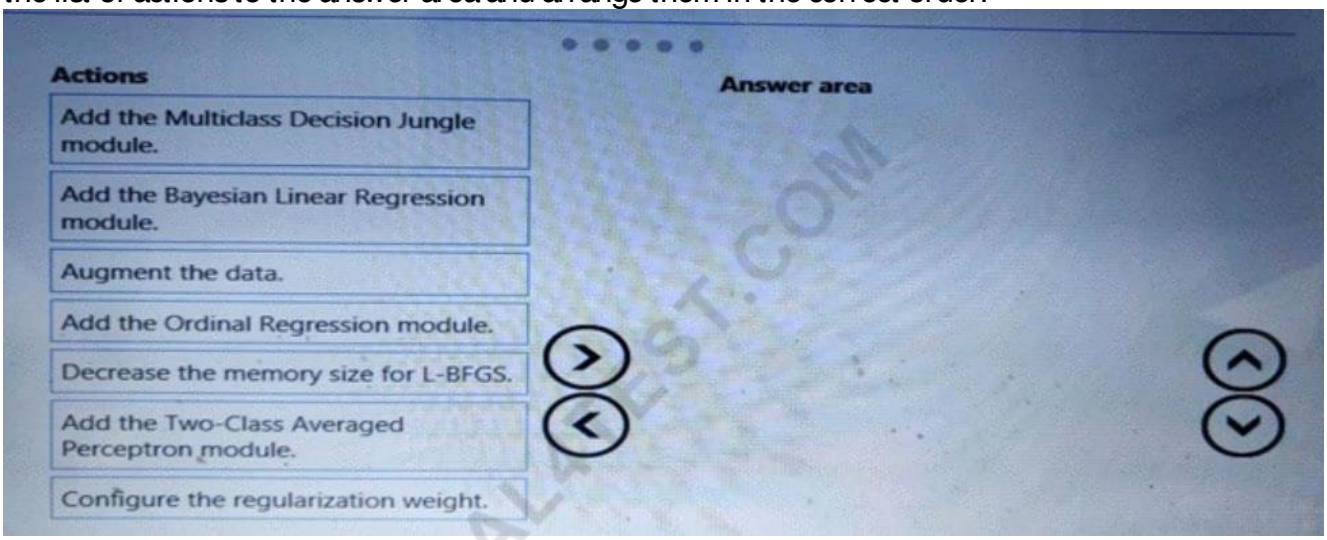
NO.56 You are developing deep learning models to analyze semi-structured, unstructured, and structured data types.
You have the following data available for model building:
* Video recordings of sporting events
* Transcripts of radio commentary about events
* Logs from related social media feeds captured during sporting events
You need to select an environment for creating the model.
Which environment should you use?
A. Azure Cognitive Services
B. Azure Data Lake Analytics
C. Azure HDInsight with Spark MLib
D. Azure Machine Learning Studio
Answer: A
Explanation
Azure Cognitive Services expand on Microsoft's evolving portfolio of machine learning APIs and enable developers to easily add cognitive features - such as emotion and video detection; facial, speech, and vision recognition; and speech and language understanding - into their applications. The goal of Azure Cognitive Services is to help developers create applications that can see, hear, speak, understand, and even begin to reason. The catalog of services within Azure Cognitive Services can be categorized into five main pillars - Vision, Speech, Language, Search, and Knowledge.
References:
https://docs.microsoft.com/en-us/azure/cognitive-services/welcome

NO.57 Your team is building a data engineering and data science development environment.
The environment must support the following requirements:
* support Python and Scala
* compose data storage, movement, and processing services into automated data pipelines
* the same tool should be used for the orchestration of both data engineering and data science
* support workload isolation and interactive workloads
* enable scaling across a cluster of machines
You need to create the environment.
What should you do?
A. Build the environment in Apache Hive for HDInsight and use Azure Data Factory for orchestration.
B. Build the environment in Azure Databricks and use Azure Data Factory for orchestration.
C. Build the environment in Apache Spark for HDInsight and use Azure Container Instances for orchestration.
D. Build the environment in Azure Databricks and use Azure Container Instances for orchestration.
Answer: B
Explanation
In Azure Databricks, we can create two different types of clusters.

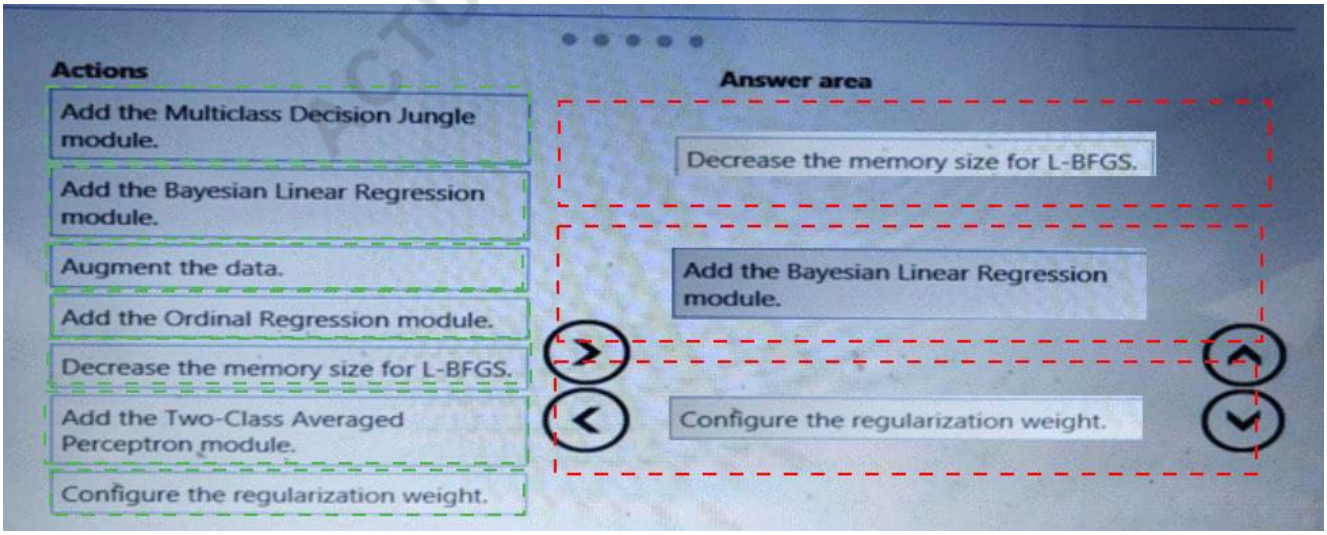
* Standard, these are the default clusters and can be used with Python, R, Scala and SQL.
* High-concurrency

Azure Databricks is fully integrated with Azure Data Factory.

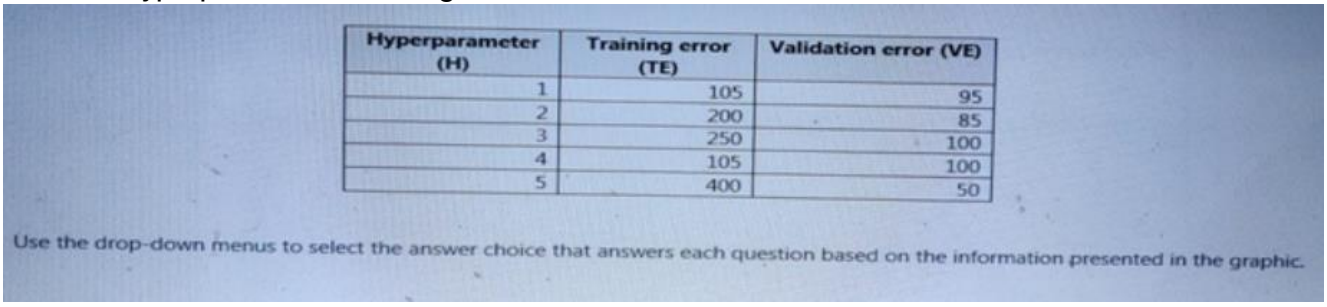
NO.58 You need to correct the model fit issue.
Which three actions should you perform in sequence? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.



Answer:

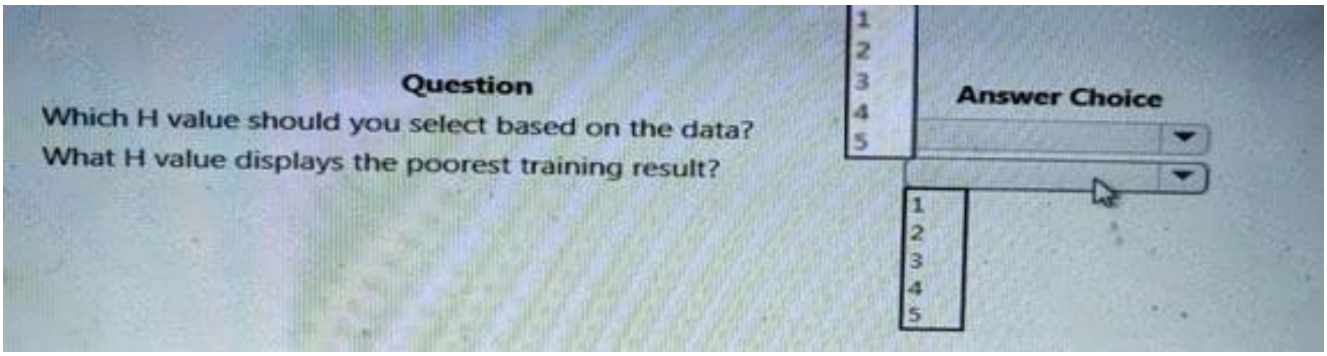


NO.59 You are tuning a hyperparameter for an algorithm. The following table shows a data set with different hyperparameter, training error, and validation errors.

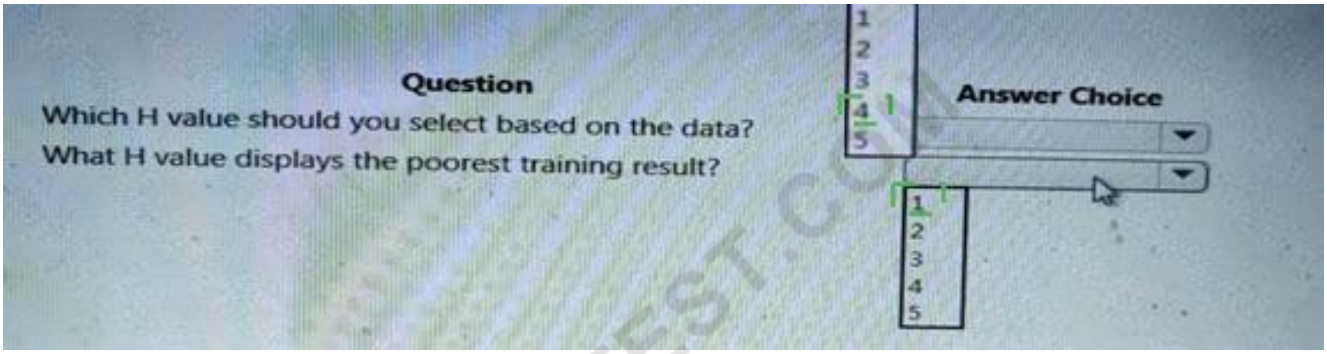


Hyperparameter (H)	Training error (TE)	Validation error (VE)
1	105	95
2	200	85
3	250	100
4	105	100
5	400	50

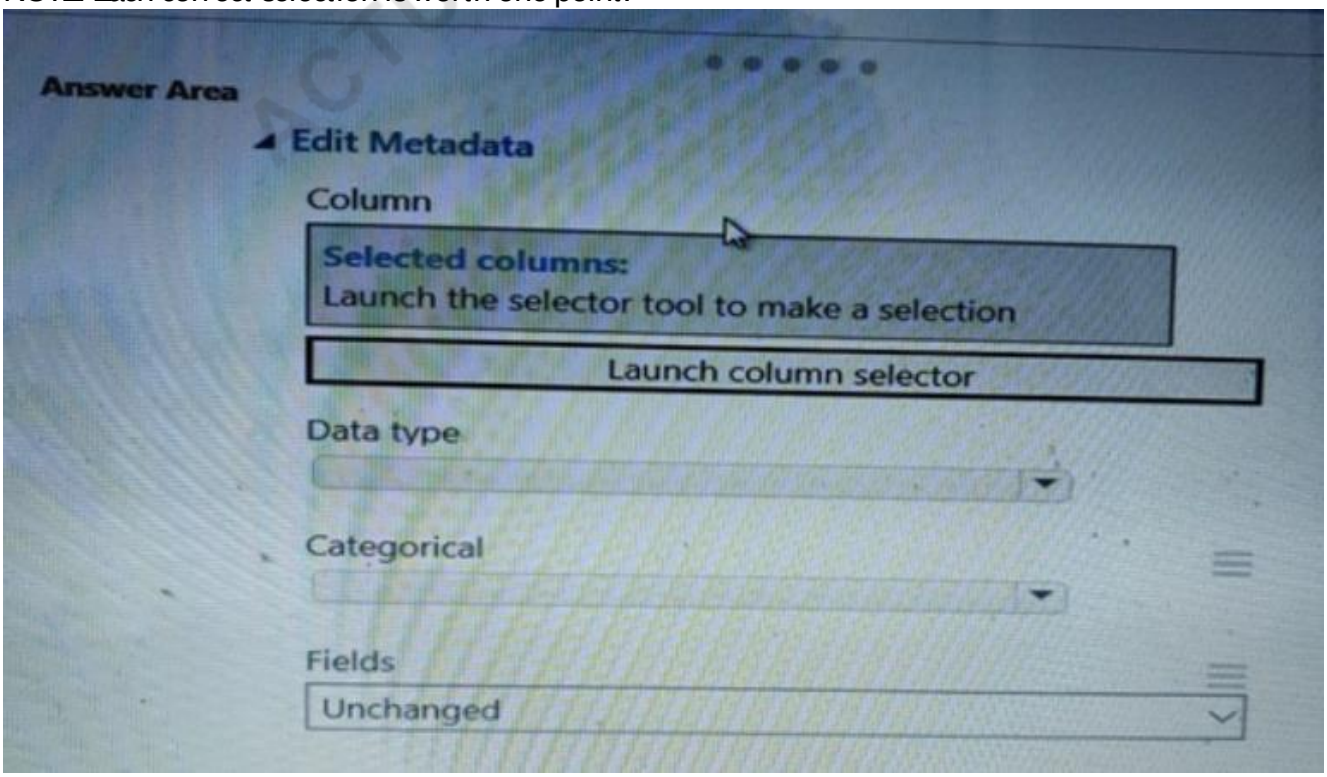
Use the drop-down menus to select the answer choice that answers each question based on the information presented in the graphic.

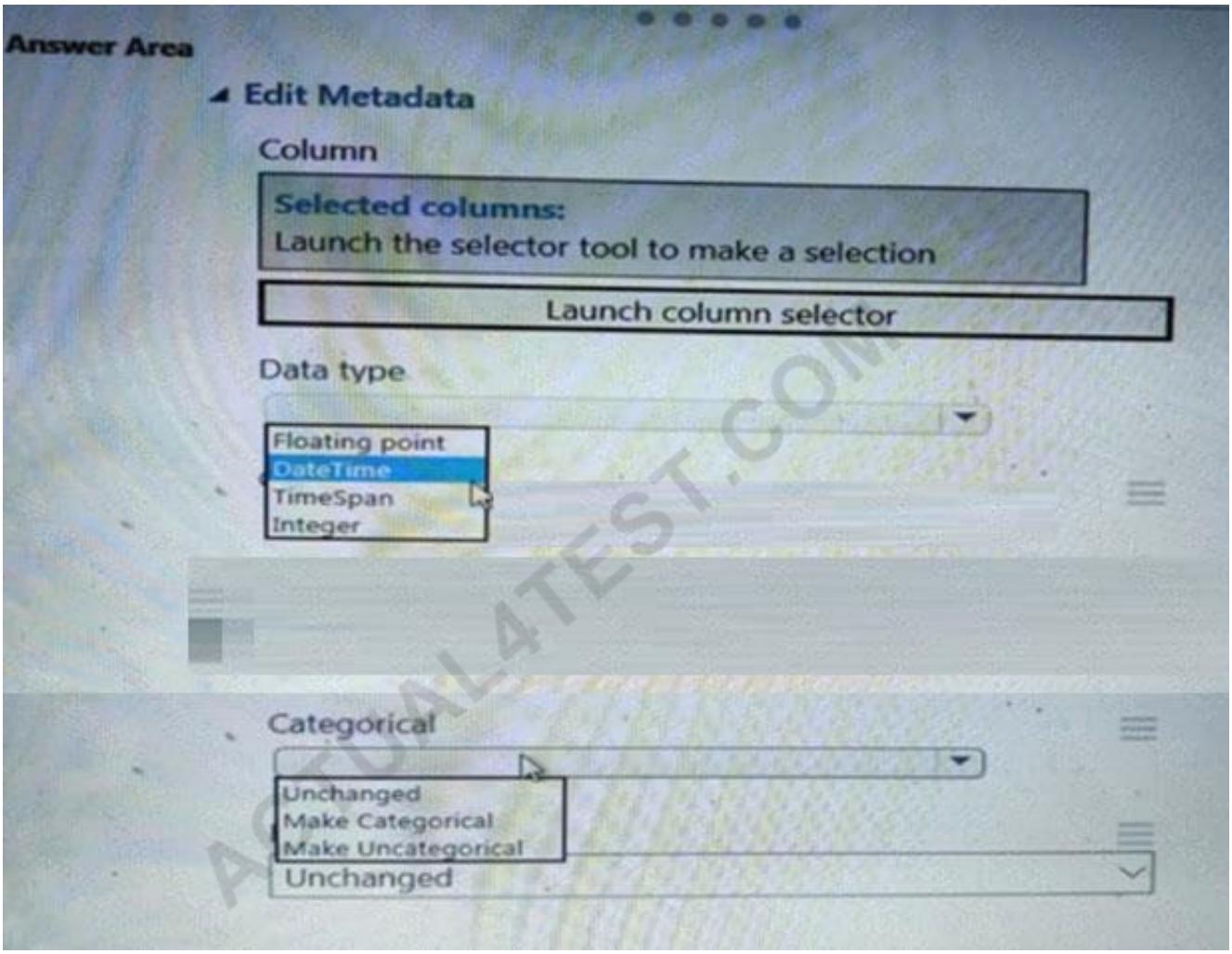


Answer:

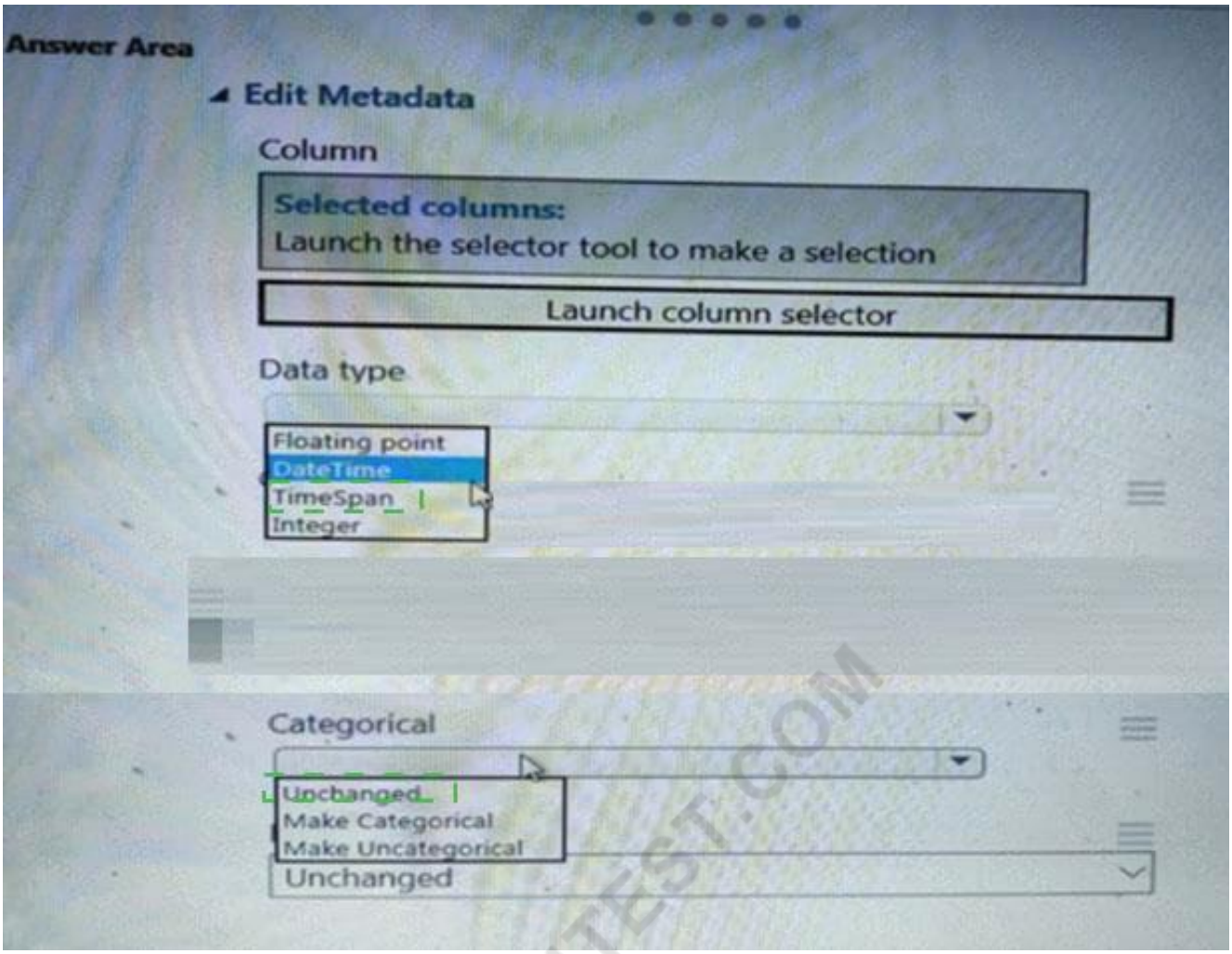


NO.60 You need to configure the Edit Metadata module so that the structure of the datasets match. Which configuration options should you select? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.

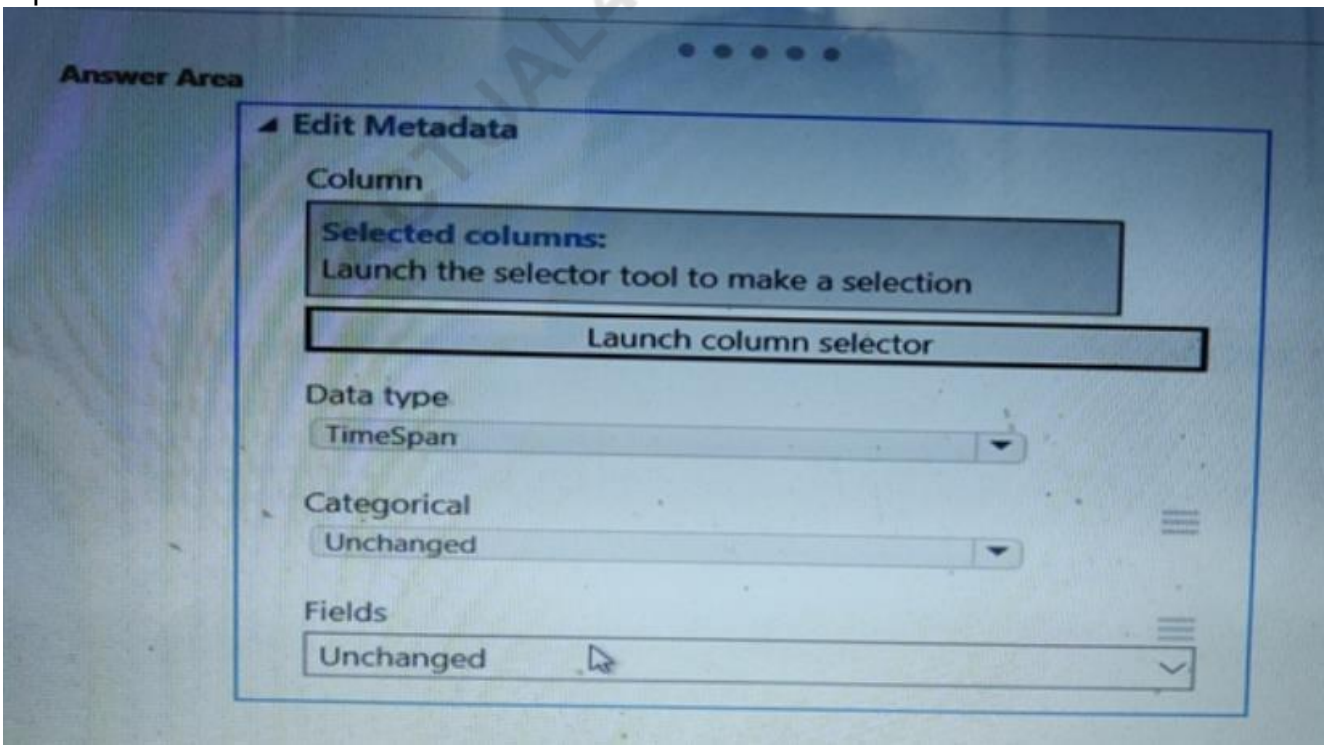



Answer Area
Edit Metadata
Column
Selected columns:
Launch the selector tool to make a selection
Launch column selector
Data type
Floating point
DateTime
TimeSpan
Integer
Categorical
Unchanged
Make Categorical
Make Uncategorical
Unchanged

Answer:



Explanation



NO.61 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 a data scientist using Azure Machine Learning Studio.
You need to normalize values to produce an output column into bins to predict a target column.
Solution: Apply a Quantiles binning mode with a PQuantile normalization.
Does the solution meet the goal?

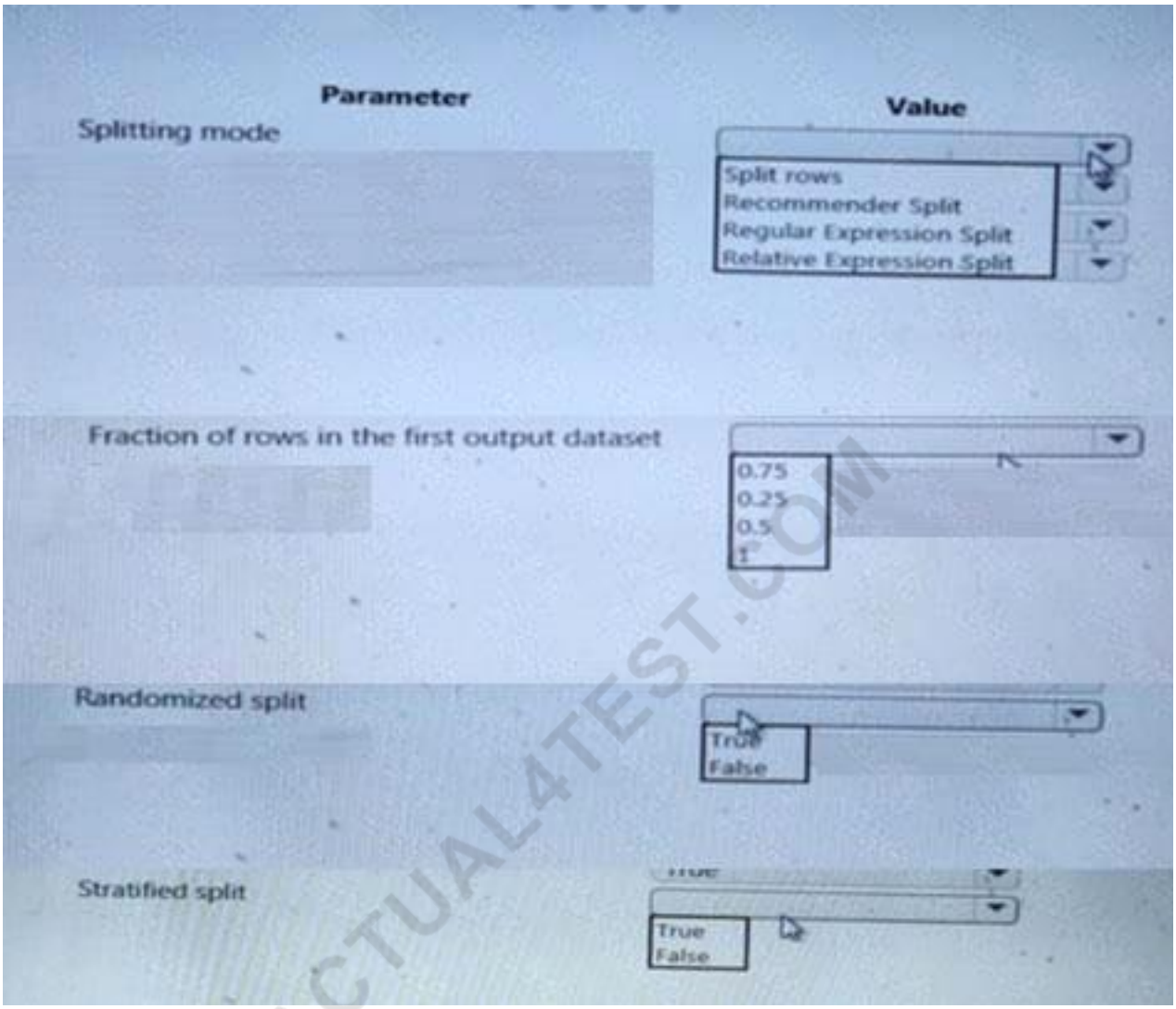
A. Yes

B. No

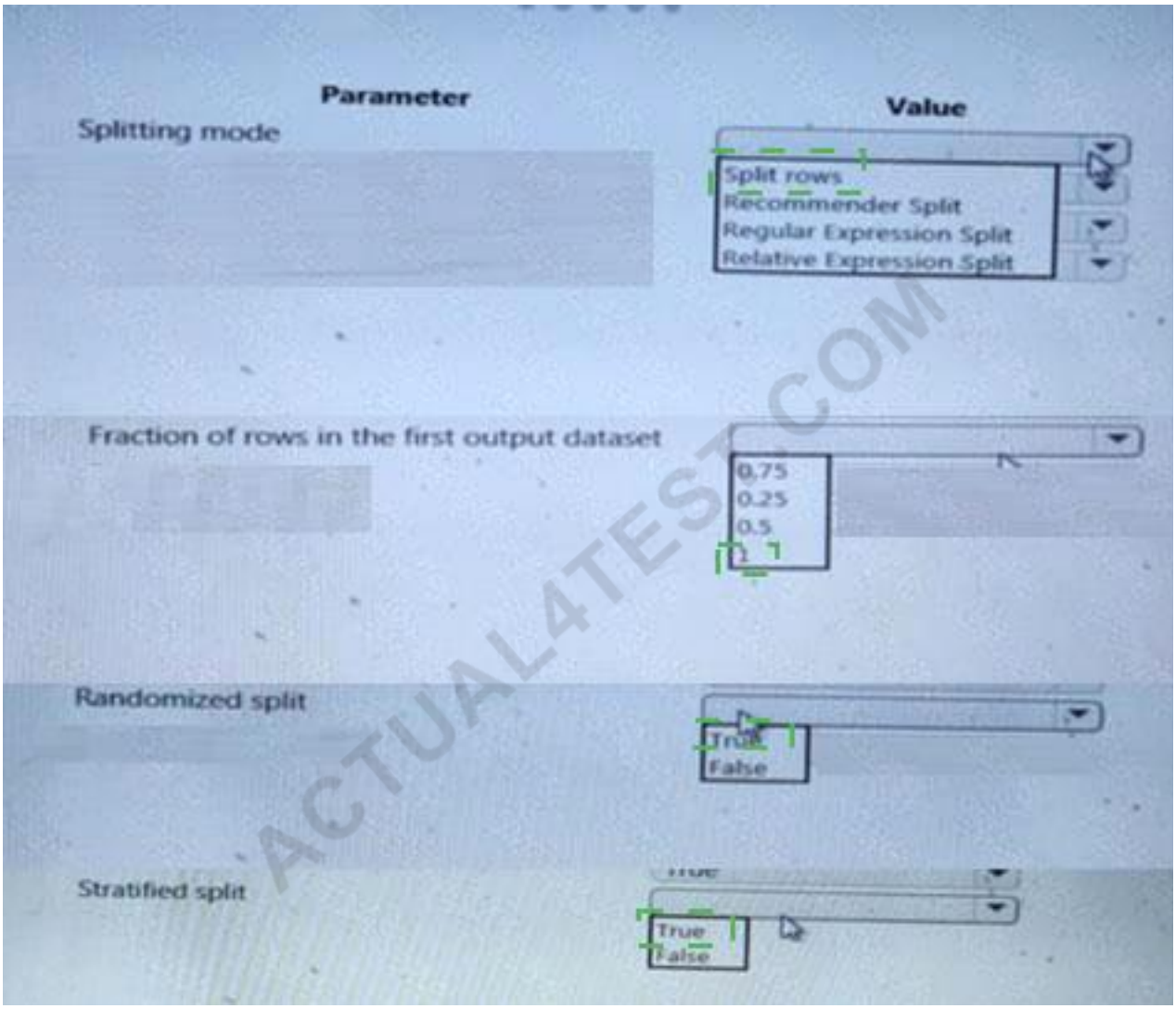
Answer: B

Explanation
Use the Entropy MDL binning mode which has a target column.
References:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/group-data-into-bins

NO.62 You are performing a classification task in Azure Machine learning Studio.
You must prepare balanced testing and training samples based on a provided data set.
Warning samples based on a provided data set.
You need to split the data with a 0.75:0.25.
Which value should you use for each parameter? To answer, select the appropriate options m the answer area.
NOTE: Each correct selection is worth one point.



Answer:



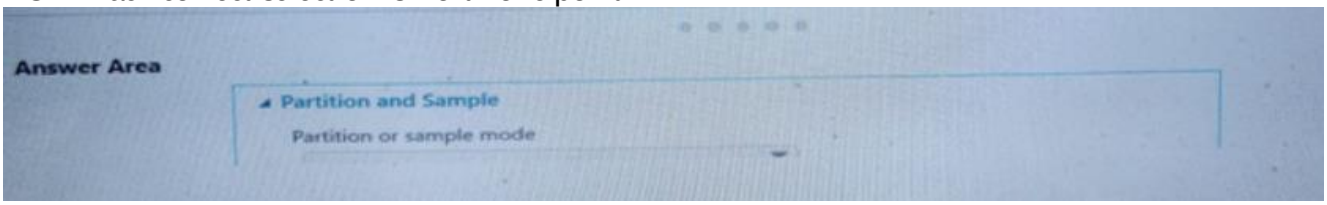
NO.63 You have a dataset contains 2,000 rows. You arc building a machine learning classification model by using Azure Machine Learning Studio. You add a Partition and Sample module to the experiment.

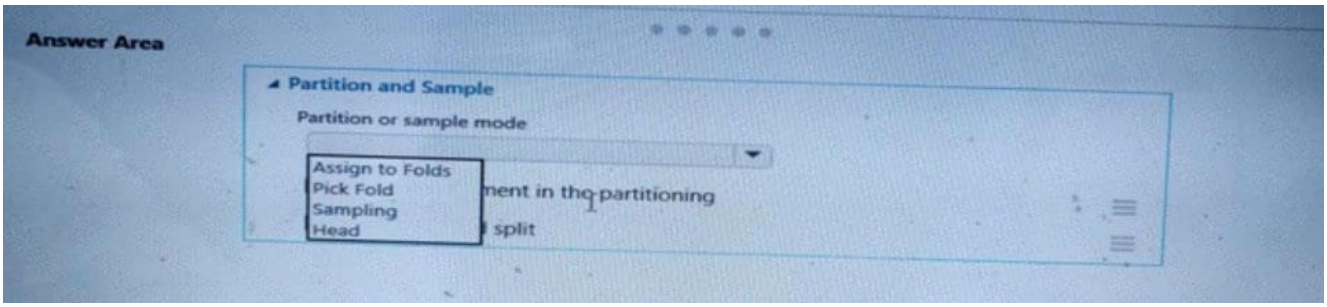
You need to configure the module. You must meet the following requirements:

* Divide the data into subsets.
* Assign the rows into folds using a round-robin method.
* Allow rows in the dataset to be reused.

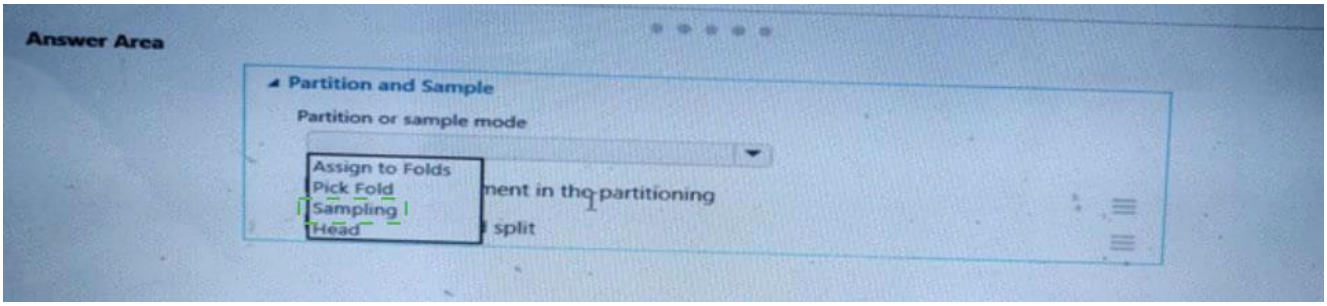
How should you configure the module? To answer select the appropriate Options m the dialog box in the answer area.

NOTE: Each correct selection is worth one point.

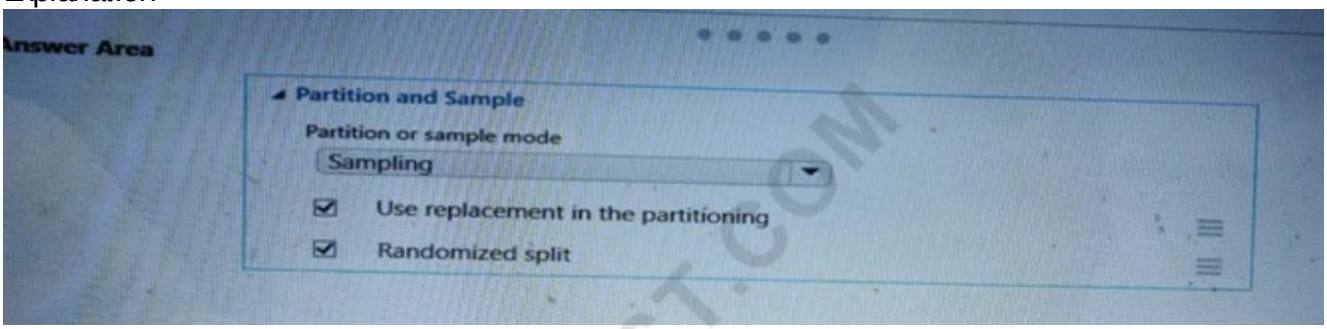




Answer:



Explanation



NO.64 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 contains 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: Calculate the column median value and use the median value as the replacement for any missing value in the column.
Does the solution meet the goal?

A. Yes

B. No

Answer: B

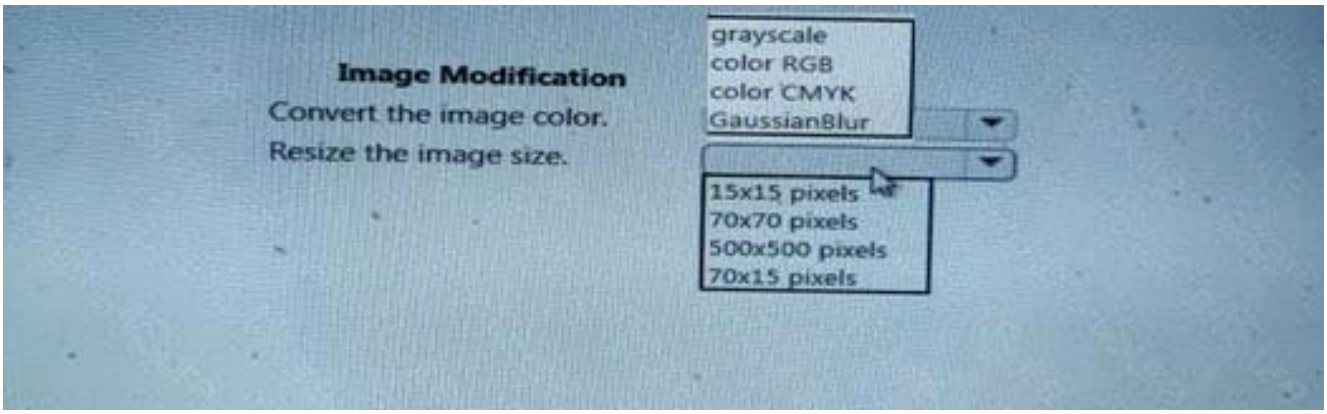
Explanation
Use the Multiple Imputation by Chained Equations (MICE) method.
References:
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3074241/
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/clean-missing-

data

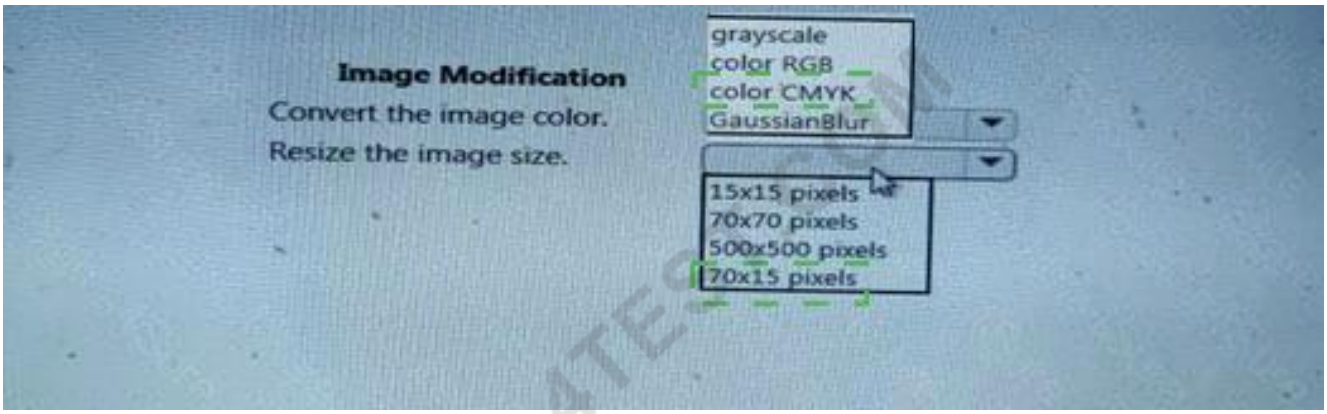
NO.65 You create a binary classification model.
You need to evaluate the model performance.
Which two metrics can you use? Each correct answer presents a complete solution.
NOTE: Each correct selection is worth one point.
A. relative absolute error
B. precision
C. accuracy
D. mean absolute error
E. coefficient of determination
Answer: B C
Explanation
The evaluation metrics available for binary classification models are: Accuracy, Precision, Recall, F1 Score, and AUC.
Note: A very natural question is: 'Out of the individuals whom the model, how many were classified correctly (TP)?' This question can be answered by looking at the Precision of the model, which is the proportion of positives that are classified correctly.
References:
https://docs.microsoft.com/en-us/azure/machine-learning/studio/evaluate-model-performance

NO.66 You plan to create a speech recognition deep learning model.
The model must support the latest version of Python.
You need to recommend a deep learning framework for speech recognition to include in the Data Science Virtual Machine (DSVM).
What should you recommend?
A. Apache Drill
B. Tensorflow
C. Rattle
D. Weka
Answer: B

NO.67 You arc I mating a deep learning model to identify cats and dogs. You have 25,000 color images.
You must meet the following requirements:
* Reduce the number of training epochs.
* Reduce the size of the neural network.
* Reduce over-fitting of the neural network.
You need to select the image modification values.
Which value should you use? To answer, select the appropriate Options in the answer area.
NOTE: Each correct selection is worth one point.



Answer:



NO.68 You are a data scientist creating a linear regression model.
You need to determine how closely the data fits the regression line.
Which metric should you review?

A. Coefficient of determination

B. Recall

C. Precision

D. Mean absolute error

E. Root Mean Square Error

Answer: A

Explanation

Coefficient of determination, often referred to as R2, represents the predictive power of the model as a value between 0 and 1. Zero means the model is random (explains nothing); 1 means there is a perfect fit. However, caution should be used in interpreting R2 values, as low values can be entirely normal and high values can be suspect.

References:

https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/evaluate-model

NO.69 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 contains 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: Remove the entire column that contains the missing data point.
Does the solution meet the goal?
A. Yes
B. No
Answer: B
Explanation
Use the Multiple Imputation by Chained Equations (MICE) method.
References:
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3074241/
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/clean-missing-data

NO.70 You create a binary classification model by using Azure Machine Learning Studio.
You must tune hyperparameters by performing a parameter sweep of the model. The parameter sweep must meet the following requirements:
* iterate all possible combinations of hyperparameters
* minimize computing resources required to perform the sweep
* You need to perform a parameter sweep of the model.
Which parameter sweep mode should you use?
A. Random sweep
B. Sweep clustering
C. Entire grid
D. Random grid
E. Random seed
Answer: D
Explanation
Maximum number of runs on random grid: This option also controls the number of iterations over a random sampling of parameter values, but the values are not generated randomly from the specified range; instead, a matrix is created of all possible combinations of parameter values and a random sampling is taken over the matrix. This method is more efficient and less prone to regional oversampling or undersampling.
If you are training a model that supports an integrated parameter sweep, you can also set a range of seed values to use and iterate over the random seeds as well. This is optional, but can be useful for avoiding bias introduced by seed selection.

NO.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 (LOCF) method to impute the missing data points.
Does the solution meet the goal?
A. Yes
B. No
Answer: B

NO.72 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 creating a model to predict the price of a student's artwork depending on the following variables: the student's length of education, degree type, and art form.
You start by creating a linear regression model.
You need to evaluate the linear regression model.
Solution: Use the following metrics: Mean Absolute Error, Root Mean Absolute Error, Relative Absolute Error, Relative Squared Error, and the Coefficient of Determination.
Does the solution meet the goal?
A. Yes
B. No
Answer: A
Explanation
The following metrics are reported for evaluating regression models. When you compare models, they are ranked by the metric you select for evaluation.
Mean absolute error (MAE) measures how close the predictions are to the actual outcomes; thus, a lower score is better.
Root mean squared error (RMSE) creates a single value that summarizes the error in the model. By squaring the difference, the metric disregards the difference between over-prediction and under-prediction.
Relative absolute error (RAE) is the relative absolute difference between expected and actual values; relative because the mean difference is divided by the arithmetic mean.
Relative squared error (RSE) similarly normalizes the total squared error of the predicted values by dividing by the total squared error of the actual values.
Mean Zero One Error (MZOE) indicates whether the prediction was correct or not. In other words: ZeroOneLoss(x,y) = 1 when x!=y; otherwise 0.
Coefficient of determination, often referred to as R2, represents the predictive power of the model as a value between 0 and 1. Zero means the model is random (explains nothing); 1 means there is a perfect fit. However, caution should be used in interpreting R2 values, as low values can be entirely normal and high values can be suspect.
AUC.
References:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/evaluate-model

NO.73 You are with a time series dataset in Azure Machine Learning Studio.
You need to split your dataset into training and testing subsets by using the Split Data module.

Which splitting mode should you use?

A. Regular Expression Split

B. Split Rows with the Randomized split parameter set to true

C. Relative Expression Split

D. Recommender Split

Answer: B

NO.74 You use the Two-Class Neural Network module in Azure Machine Learning Studio to build a binary classification model. You use the Tune Model Hyperparameters module to tune accuracy for the model.

You need to select the hyperparameters that should be tuned using the Tune Model Hyperparameters module.

Which two hyperparameters should you use? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

A. Number of hidden nodes

B. Learning Rate

C. The type of the normalizer

D. Number of learning iterations

E. Hidden layer specification

Answer: D E

Explanation

D: For Number of learning iterations, specify the maximum number of times the algorithm should process the training cases.

E: For Hidden layer specification, select the type of network architecture to create.

Between the input and output layers you can insert multiple hidden layers. Most predictive tasks can be accomplished easily with only one or a few hidden layers.

References:

https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/two-class-neural-network

NO.75 You are performing a filter based feature selection for a dataset 10 build a multi class classifies by using Azure Machine Learning Studio.

The dataset contains categorical features that are highly correlated to the output label column.

You need to select the appropriate feature scoring statistical method to identify the key predictors.

Which method should you use?

A. Chi-squared

B. Spearman correlation

C. Kendall correlation

D. Person correlation

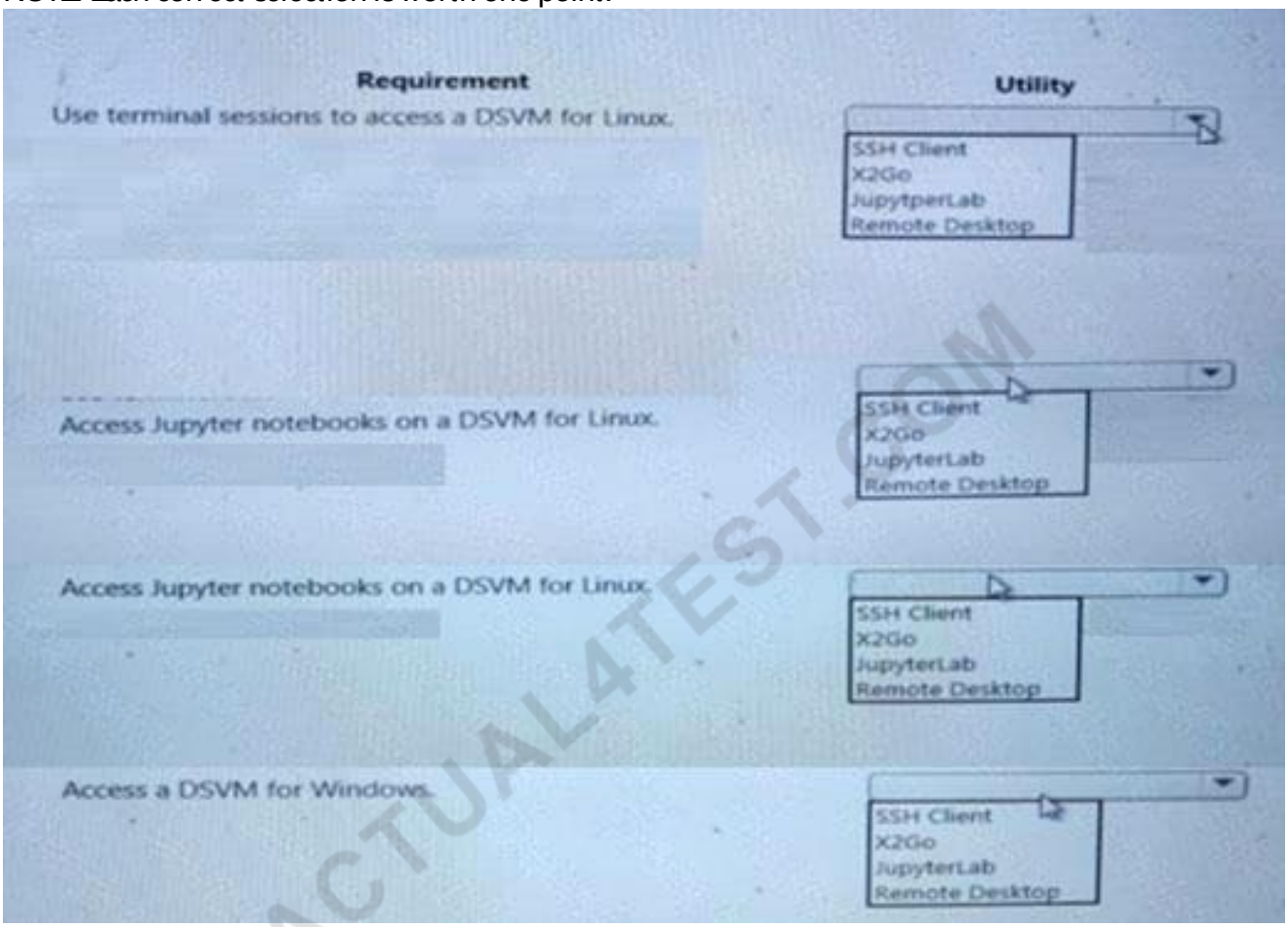
Answer: A

NO.76 You use Data Science Virtual Machines (DSVMs) for Windows and Linux in Azure.

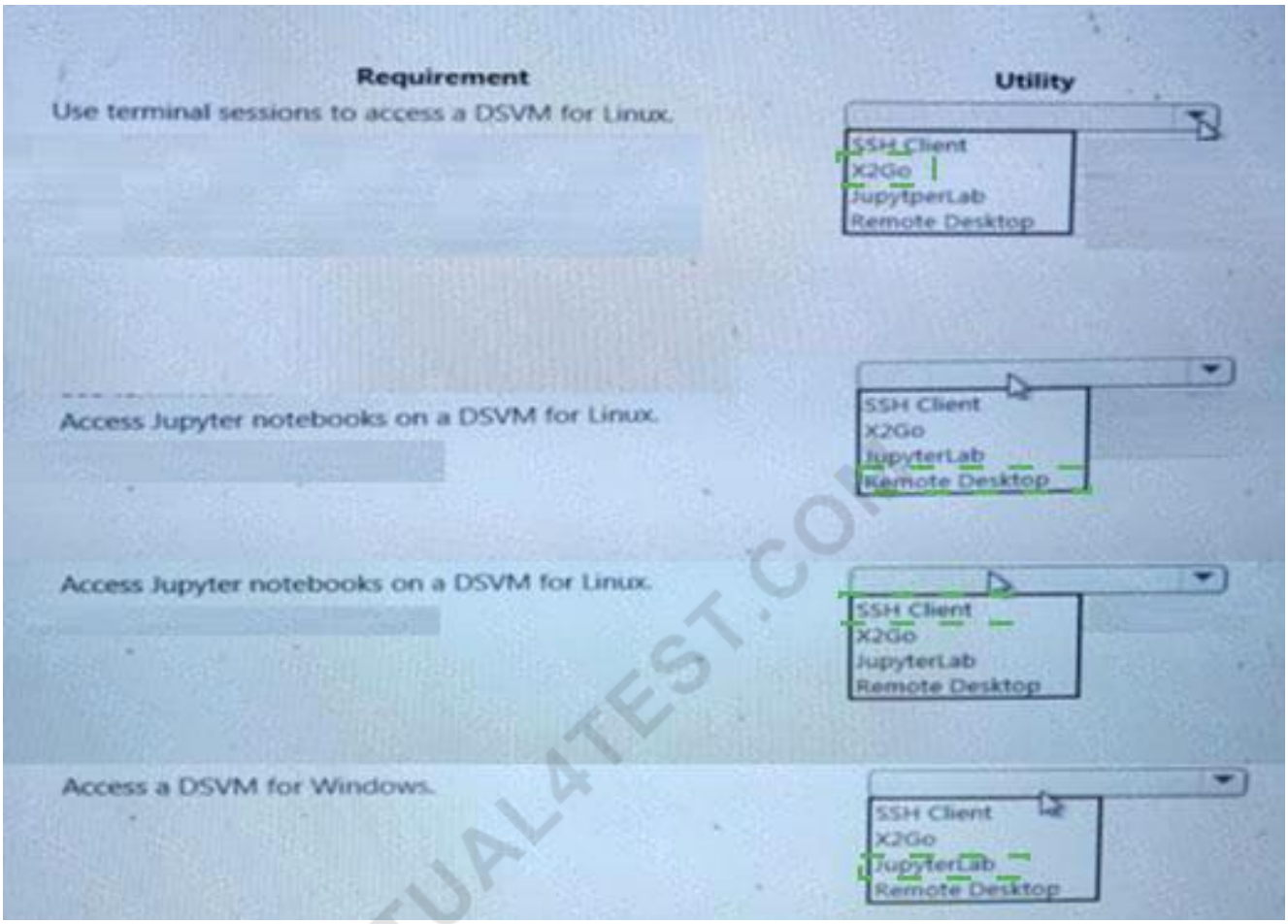
You need to access the DSVMs.

Which utilities should you use? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.



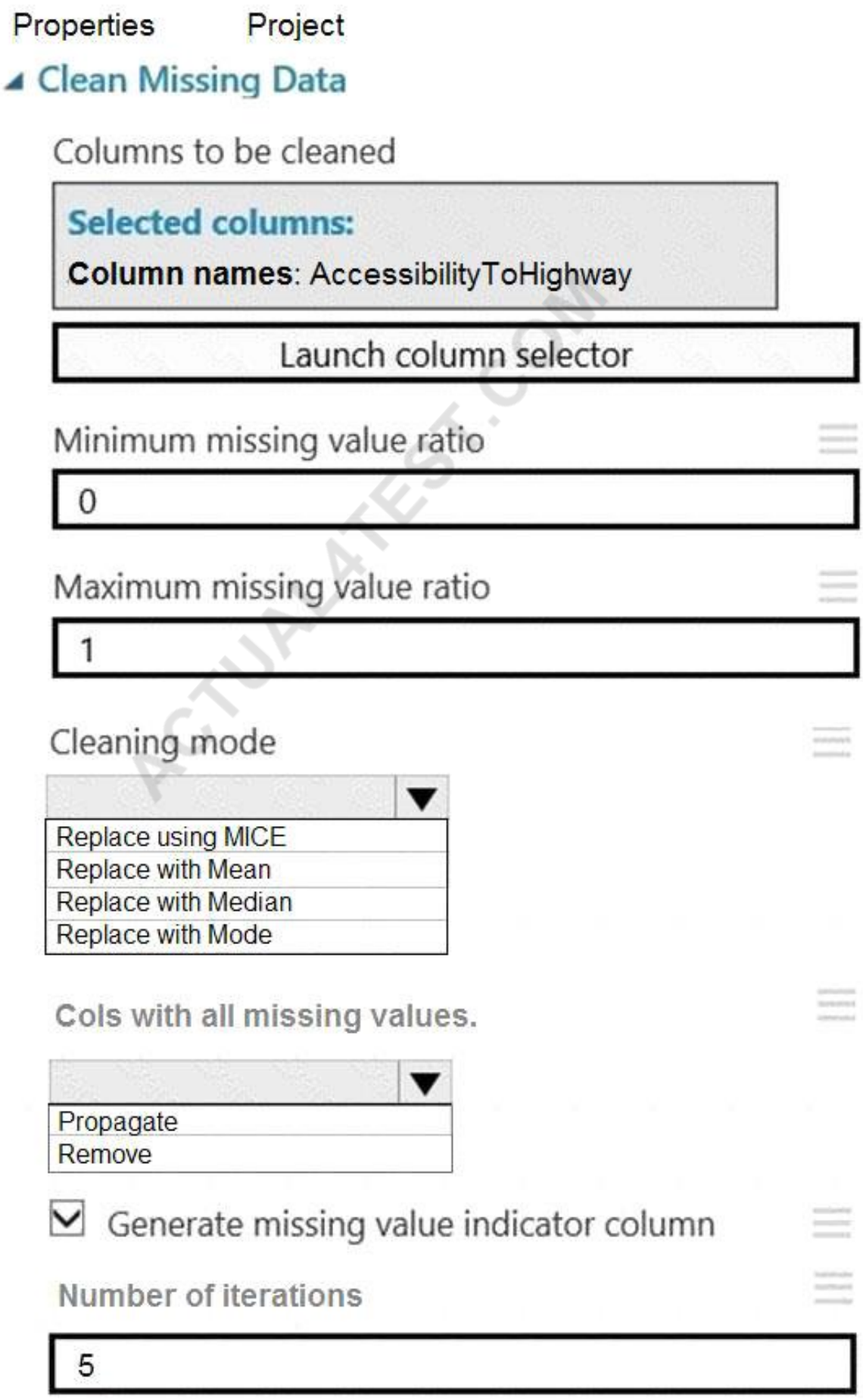
Answer:



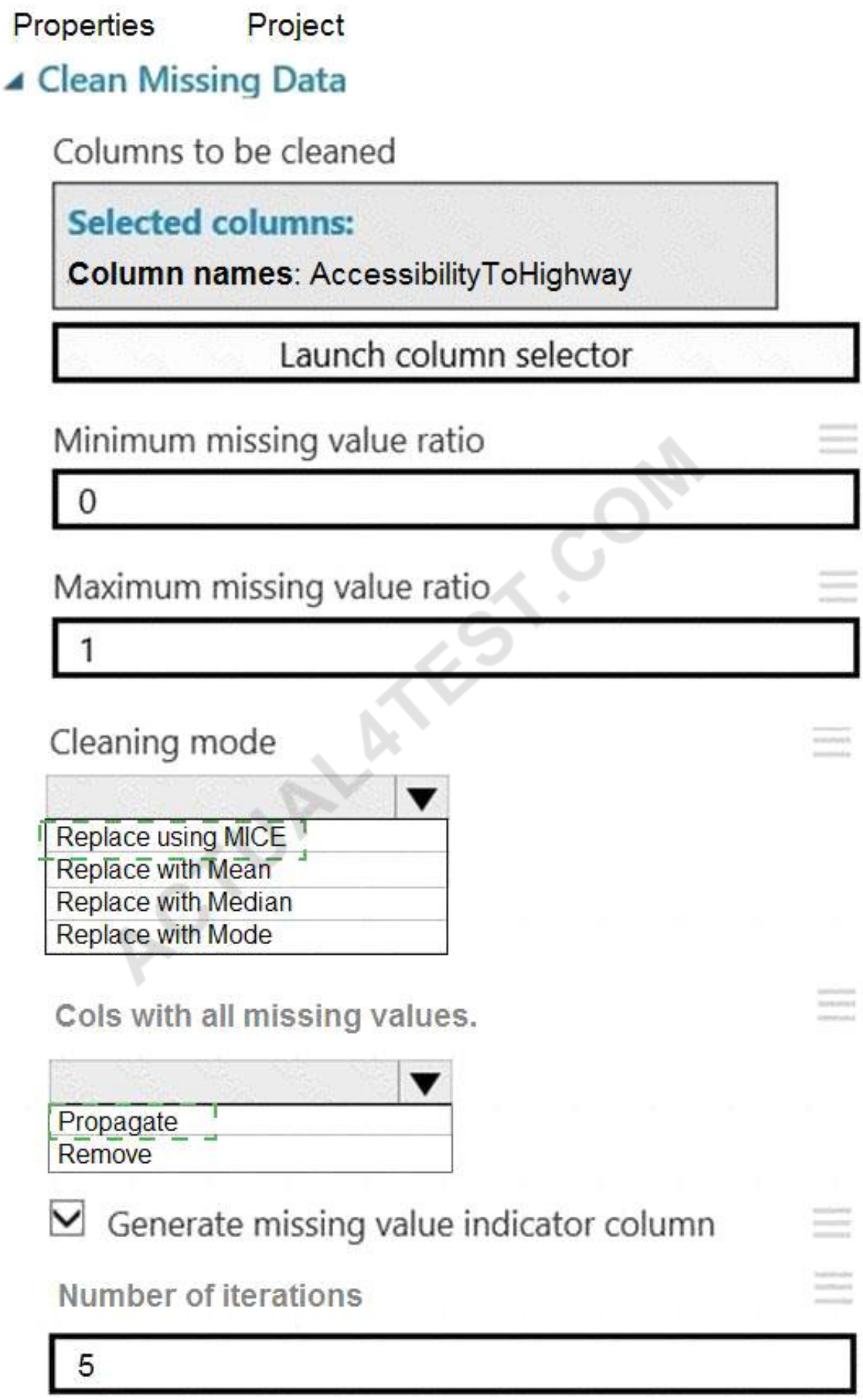
NO.77 Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution. After you answer 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 creating a model to predict the price of a student's artwork depending on the following variables: the student's length of education, degree type, and art form.
You start by creating a linear regression model.
You need to evaluate the linear regression model.
Solution: Use the following metrics: Relative Squared Error, Coefficient of Determination, Accuracy, Precision, Recall, F1 score, and AUC.
Does the solution meet the goal?
A. Yes
B. No
Answer: B
Explanation
Relative Squared Error, Coefficient of Determination are good metrics to evaluate the linear regression model, but the others are metrics for classification models.
References:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/evaluate-model

NO.78 You need to replace the missing data in the AccessibilityToHighway columns.

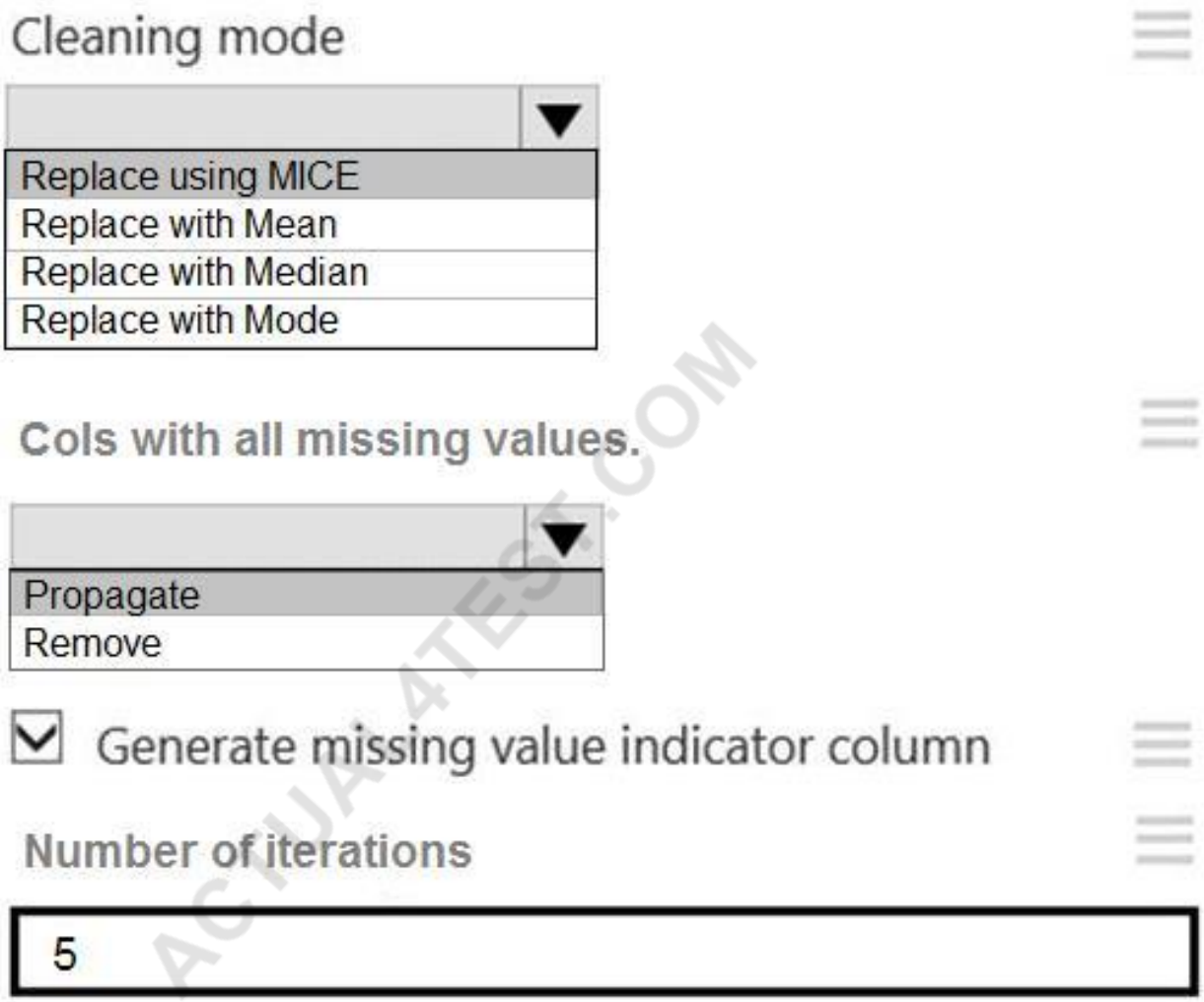
How should you configure the Clean Missing Data module? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.



Answer:



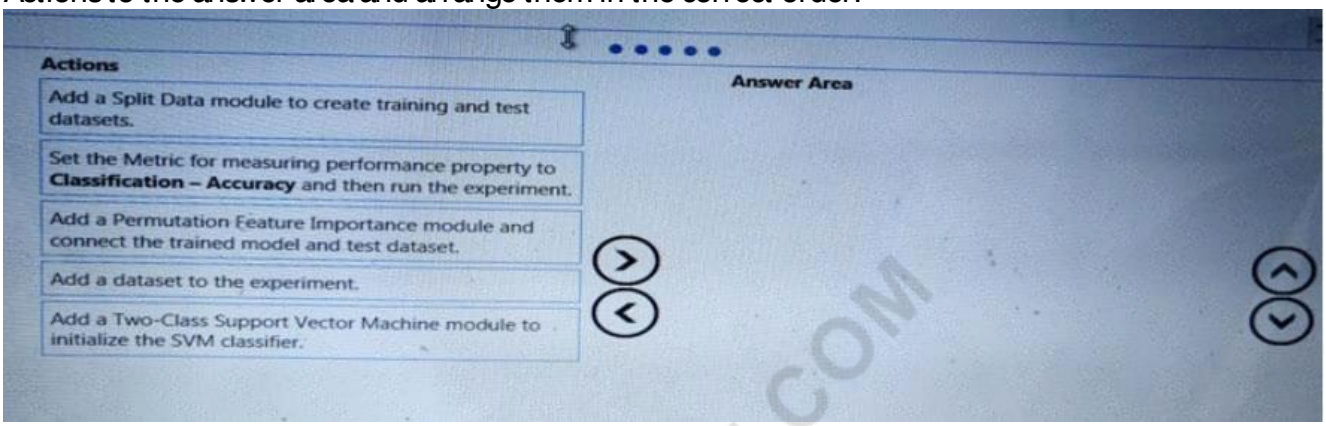
Explanation



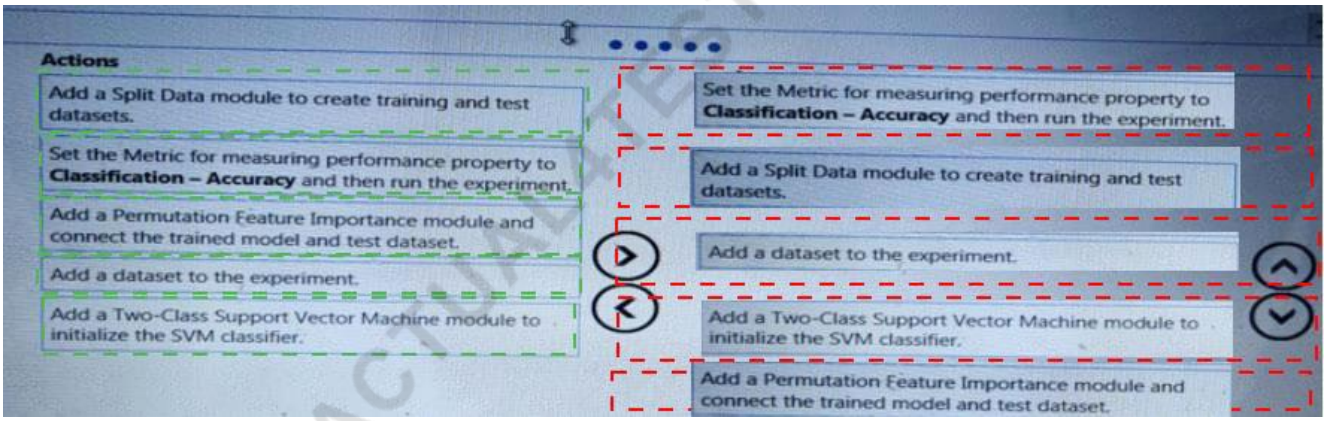
Box 1: Replace using MICE
Replace using MICE: For each missing value, this option assigns a new value, which is calculated by using a method described in the statistical literature as "Multivariate Imputation using Chained Equations" or
"Multiple Imputation by Chained Equations". With a multiple imputation method, each variable with missing data is modeled conditionally using the other variables in the data before filling in the missing values.
Scenario: The AccessibilityToHighway column in both datasets contains missing values. The missing data must be replaced with new data so that it is modeled conditionally using the other variables in the data before filling in the missing values.
Box 2: Propagate
Cols with all missing values indicate if columns of all missing values should be preserved in the output.
References:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/clean-missing-data

NO.79 YOU have a data-set that contains over 150 features. You use the dataset to train a Support Vector Machine (SVM) binary classifirer.
You need to use the Permutation Feature Importance module in Azure Machine Learning Studio to

compute a set of feature importance scores for the dataset.
In which order should you perform the actions? To answer move al actions from from the list of Actions to the answer area and arrange them in the correct order.



Answer:



NO.80 You are solving a classification task.
You must evaluate your model on a limited data sample by using k-fold cross validation. You start by configuring a k parameter as the number of splits.
You need to configure the k parameter for the cross-validation.
Which value should you use?

A. k=0.5

B. k=0

C. k=5

D. k=1

Answer: C

Explanation
Leave One Out (LOO) cross-validation
Setting K = n (the number of observations) yields n-fold and is called leave-one out cross-validation (LOO), a special case of the K-fold approach.
LOO CV is sometimes useful but typically doesn't shake up the data enough. The estimates from each fold are highly correlated and hence their average can have high variance.
This is why the usual choice is K=5 or 10. It provides a good compromise for the bias-variance tradeoff.

NO.81 You are building a machine learning model for translating English language textual content

into French language textual content.
You need to build and train the machine learning model to learn the sequence of the textual content.
Which type of neural network should you use?

A. Multilayer Perceptions (MLPs)

B. Convolutional Neural Networks (CNNs)

C. Recurrent Neural Networks (RNNs)

D. Generative Adversarial Networks (GANs)

Answer: C

Explanation
To translate a corpus of English text to French, we need to build a recurrent neural network (RNN).
Note: RNNs are designed to take sequences of text as inputs or return sequences of text as outputs, or both.
They're called recurrent because the network's hidden layers have a loop in which the output and cell state from each time step become inputs at the next time step. This recurrence serves as a form of memory. It allows contextual information to flow through the network so that relevant outputs from previous time steps can be applied to network operations at the current time step.
References:
https://towardsdatascience.com/language-translation-with-rnns-d84d43b40571

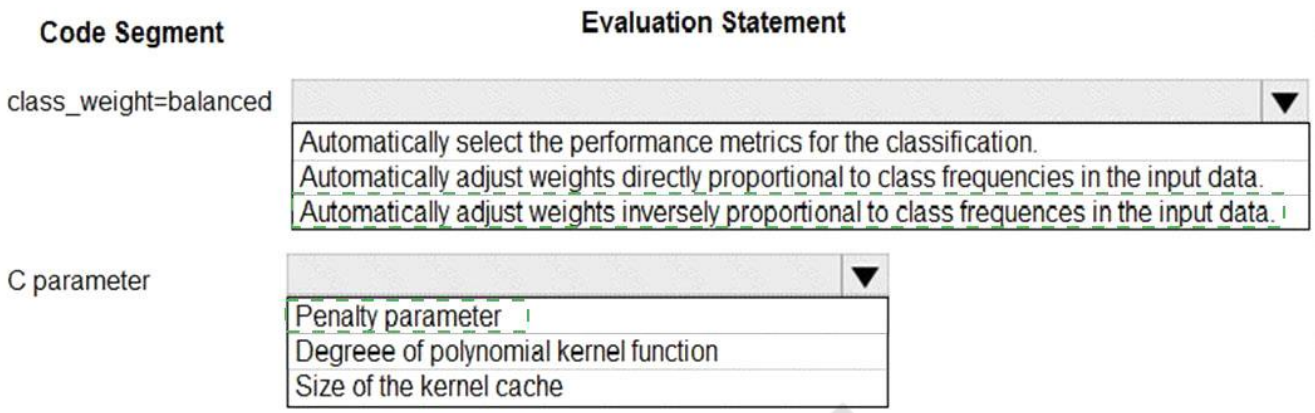
NO.82 You are using C-Support Vector classification to do a multi-class classification with an unbalanced training dataset. The C-Support Vector classification using Python code shown below:

```
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)
```

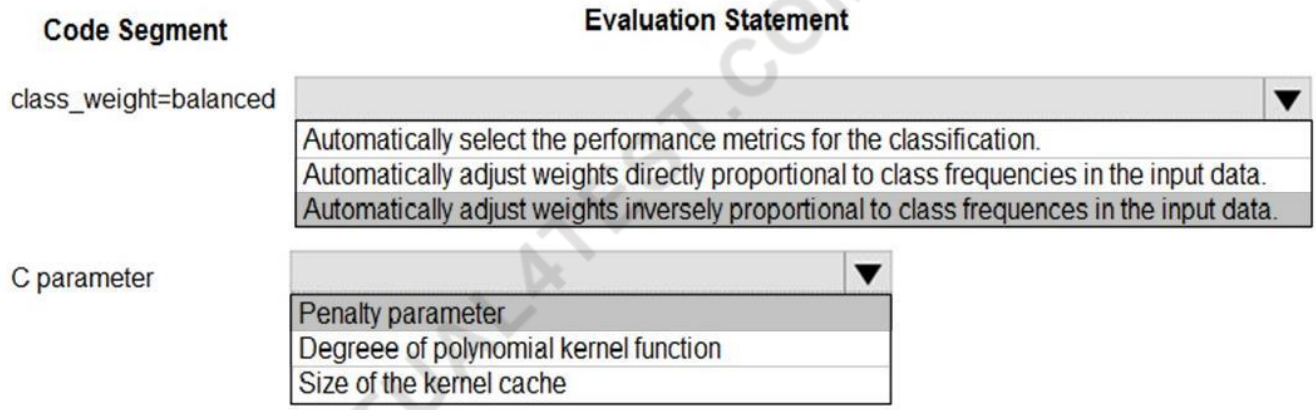
You need to evaluate the C-Support Vector classification code.
Which evaluation statement should you use? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.

Code Segment	Evaluation Statement
class_weight=balanced	Automatically select the performance metrics for the classification. Automatically adjust weights directly proportional to class frequencies in the input data. Automatically adjust weights inversely proportional to class frequences in the input data.
C parameter	Penalty parameter Degreee of polynomial kernel function Size of the kernel cache

Answer:

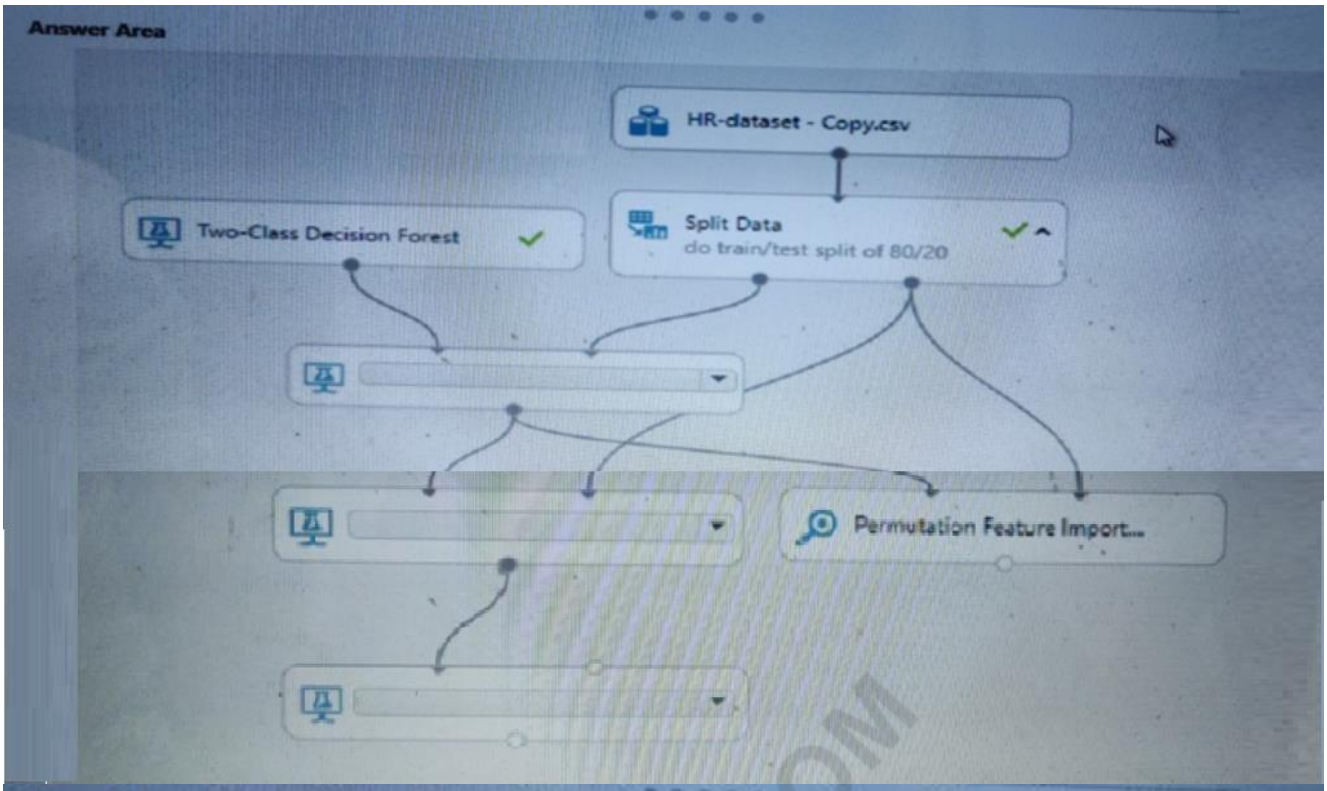


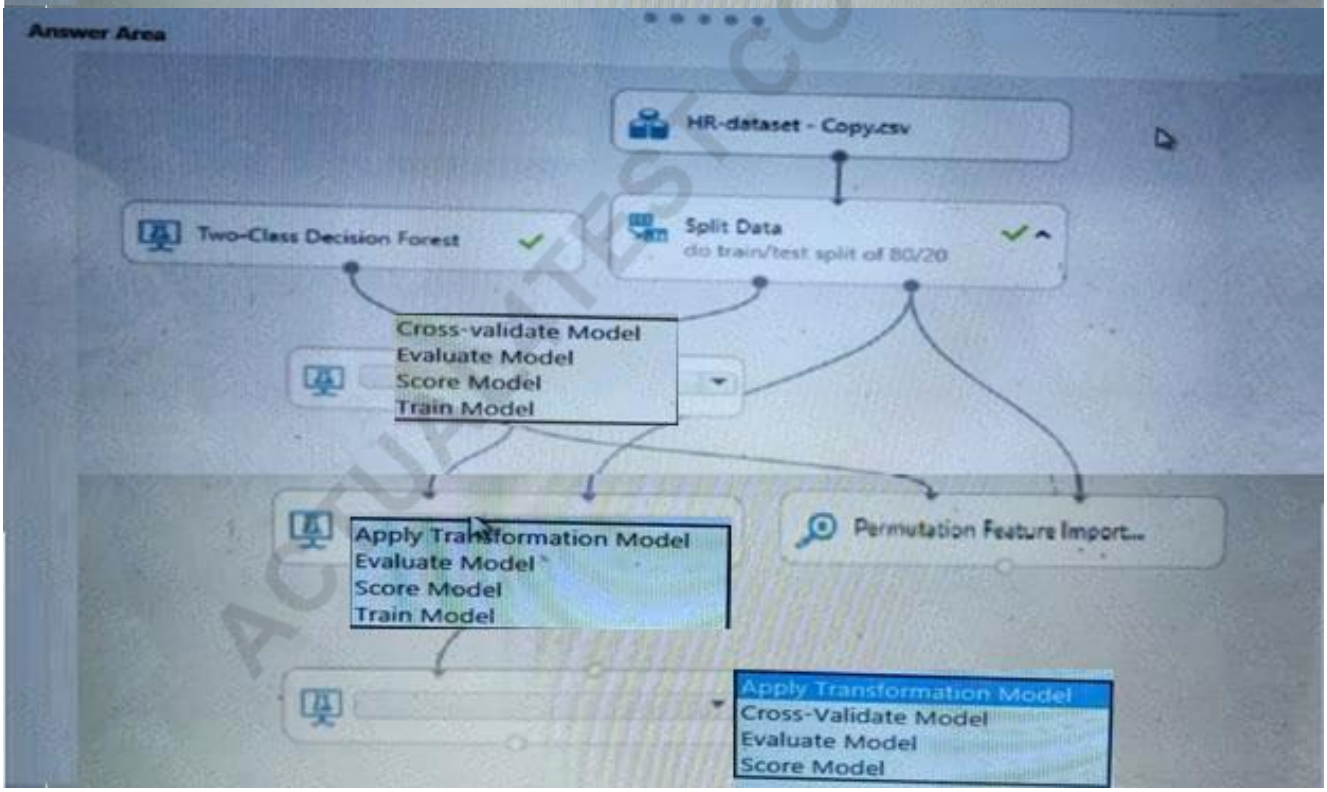
Explanation



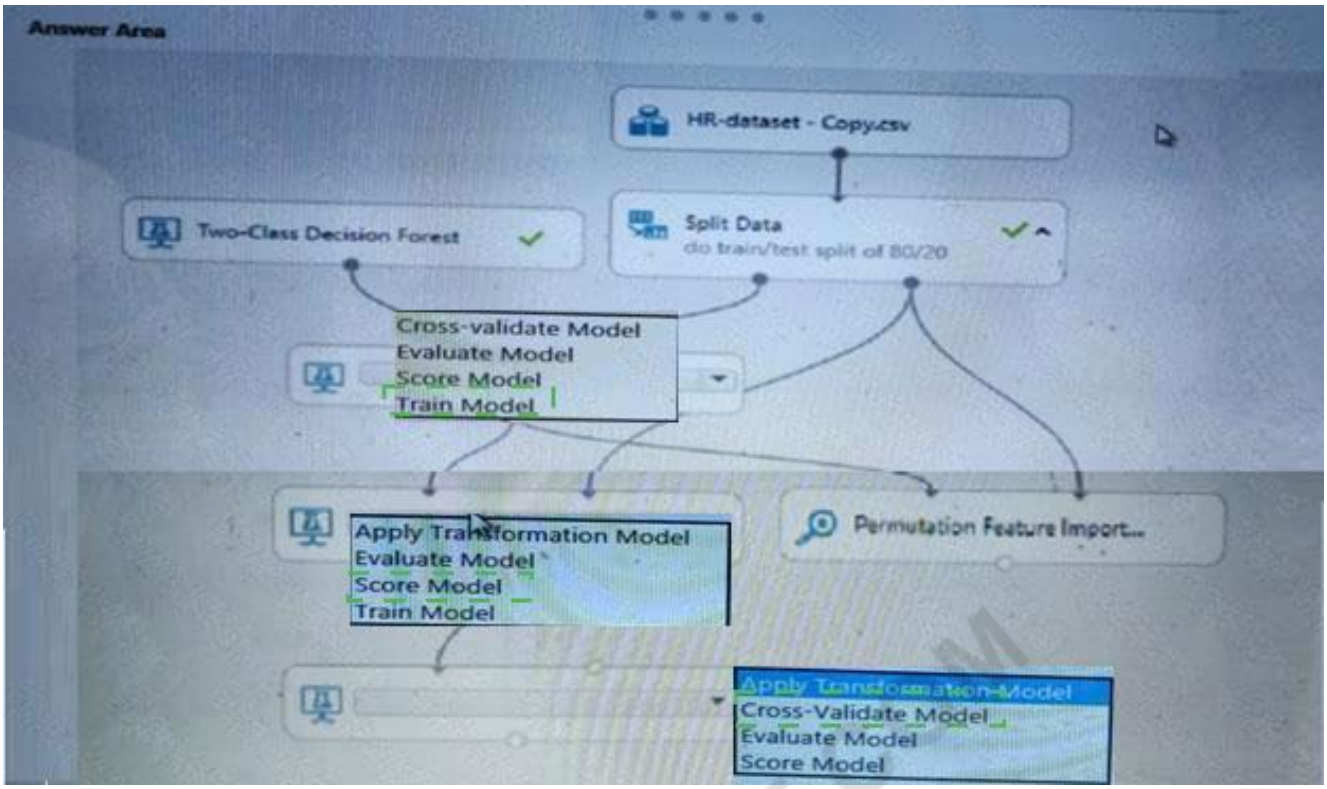
Box 1: Automatically adjust weights inversely proportional to class frequencies in the input data The "balanced" mode uses the values of y to automatically adjust weights inversely proportional to class frequencies in the input data as n_samples / (n_classes * np.bincount(y)).
Box 2: Penalty parameter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

NO.83 You create a binary classification model using Azure Machine Learning Studio.
You must use a Receiver Operating Characteristic (ROC) curve and an F1 score to evaluate the model.
You need to create the required business metrics.
How should you complete the experiment? To answer, select the appropriate options in the dialog box in the answer area.
NOTE: Each correct selection is worth one point.


Answer Area
HR-dataset - Copy.csv
Two-Class Decision Forest
Split Data
do train/test split of 80/20
Permutation Feature Import...


Answer Area
HR-dataset - Copy.csv
Two-Class Decision Forest
Split Data
do train/test split of 80/20
Cross-validate Model
Evaluate Model
Score Model
Train Model
Permutation Feature Import...
Apply Transformation Model
Evaluate Model
Score Model
Train Model
Apply Transformation Model
Cross-Validate Model
Evaluate Model
Score Model

Answer:



NO.84 You are implementing a machine learning model to predict stock prices.
The model uses a PostgreSQL database and requires GPU processing.
You need to create a virtual machine that is pre-configured with the required tools.
What should you do?

A. Create a Data Science Virtual Machine (DSVM) Windows edition.

B. Create a Geo AI Data Science Virtual Machine (Geo-DSVM) Windows edition.

C. Create a Deep Learning Virtual Machine (DLVM) Linux edition.

D. Create a Deep Learning Virtual Machine (DLVM) Windows edition.

E. Create a Data Science Virtual Machine (DSVM) Linux edition.

Answer: E

NO.85 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 creating a new experiment in Azure Learning learning Studio.
One class has a much smaller number of observations than the other classes in the training You need to select an appropriate data sampling strategy to compensate for the class imbalance.
Solution: You use the Synthetic Minority Oversampling Technique (SMOTE) sampling mode.
Does the solution meet the goal?

A. Yes

B. No

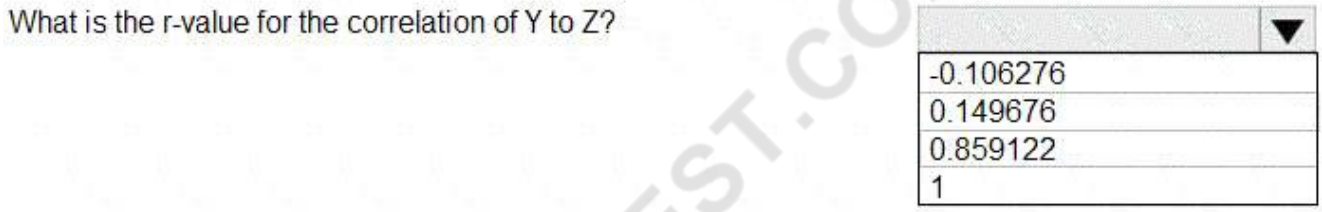
Answer: A

NO.86 You are analyzing a dataset by using Azure Machine Learning Studio.
YOU need to generate a statistical summary that contains the p value and the unique value count for each feature column.
Which two modules can you users? Each correct answer presents a complete solution.
NOTE: Each correct selection is worth one point.

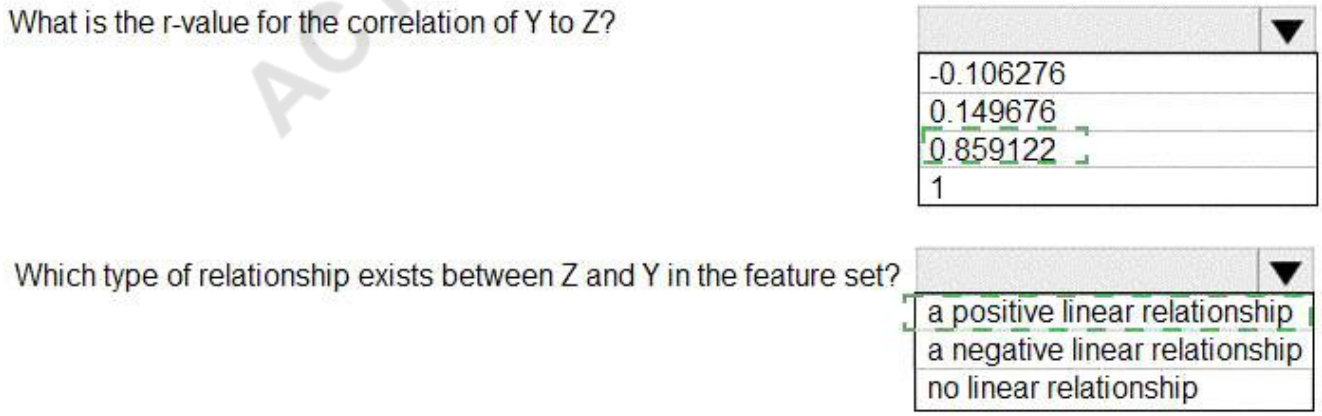
A. Execute Python Script
B. Export Count Table
C. Convert to Indicator Values
D. Summarize Data
E. Compute linear Correlation

Answer: B E

NO.87 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:
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.

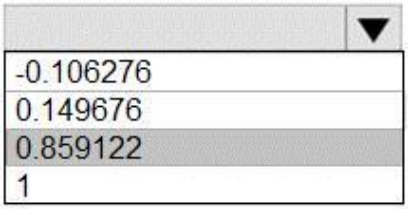


Answer:

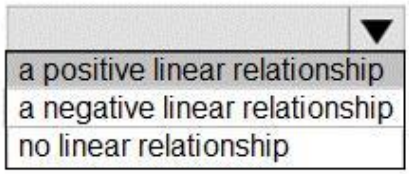


Explanation

What is the r-value for the correlation of Y to Z?



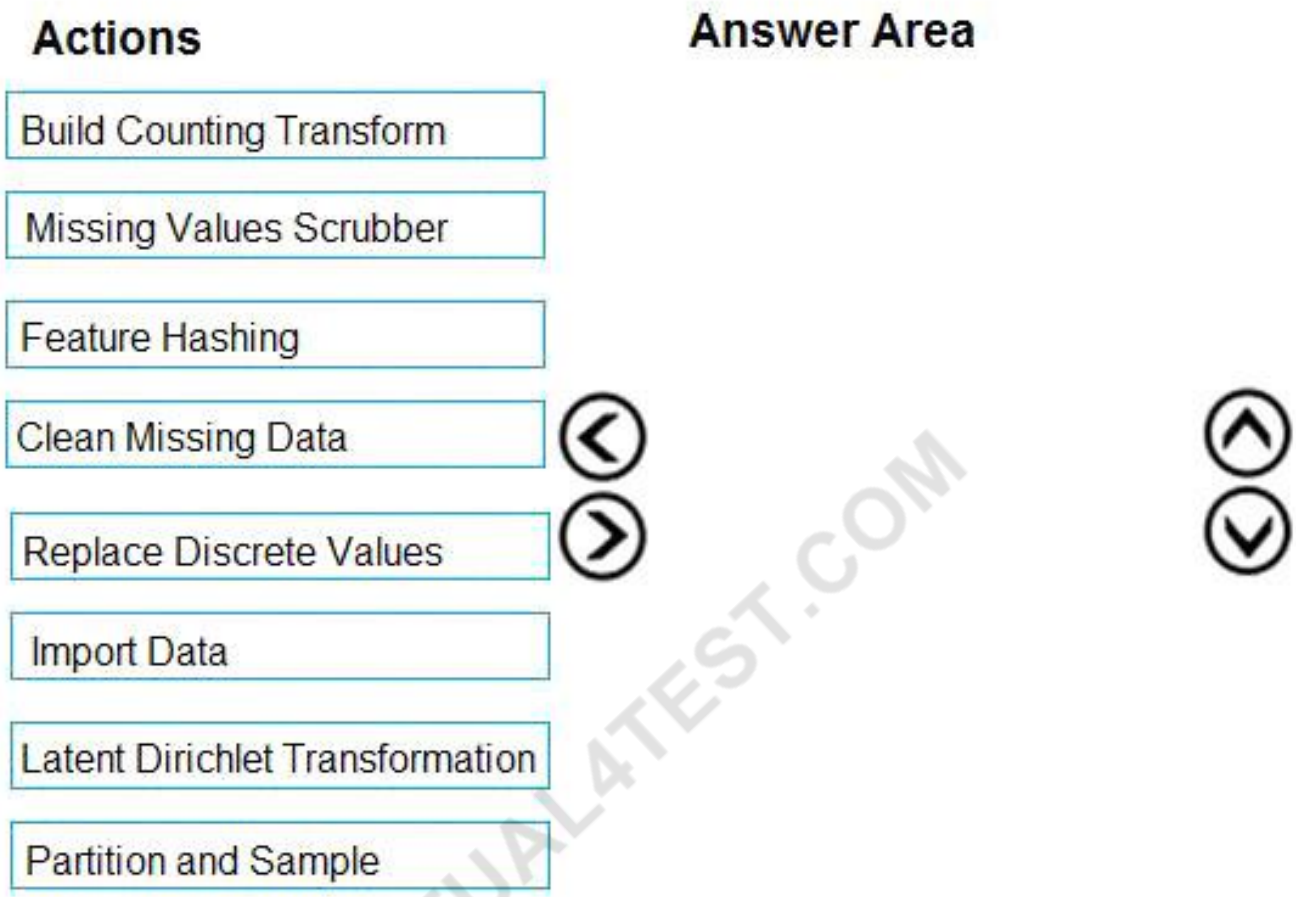
Which type of relationship exists between Z and Y in the feature set?



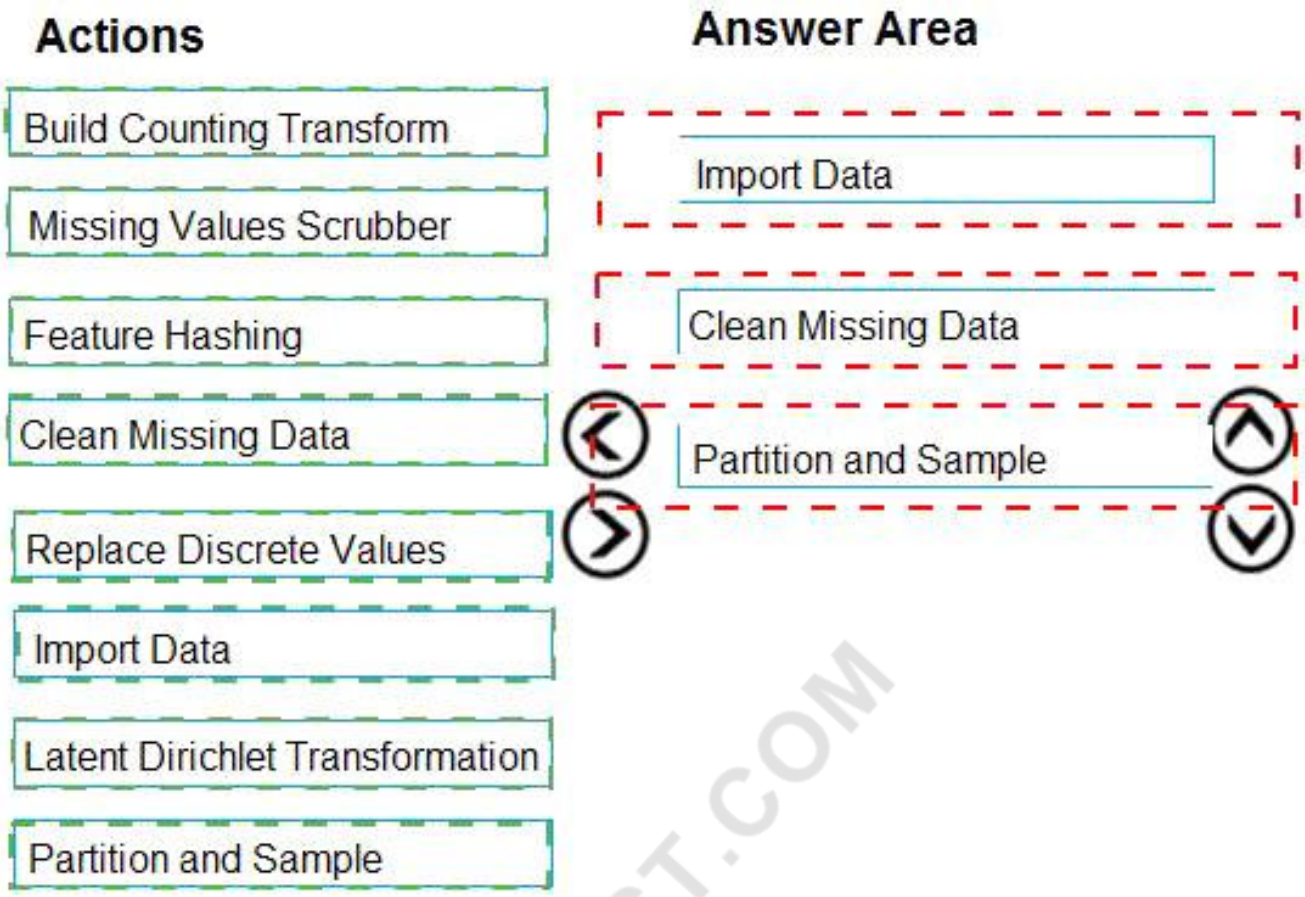
Box 1: 0.859122
Box 2: a positively linear relationship
+1 indicates a strong positive linear relationship
-1 indicates a strong negative linear correlation
0 denotes no linear relationship between the two variables.
References:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/compute-linear-correlation

NO.88 You are evaluating, a completed binary classification machine learning model.
You need to use the precision as the evaluation metric.
Which visualization should you use?
A. Binary classification confusion matrix
B. box plot
C. Gradient descent
D. coefficient of determination
Answer: B

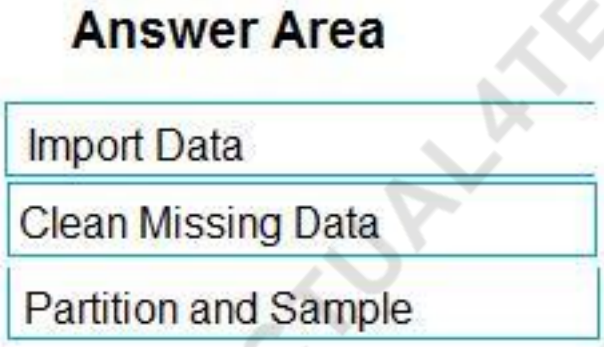
NO.89 You are creating an experiment by using Azure Machine Learning Studio.
You must divide the data into four subsets for evaluation. There is a high degree of missing values in the data.
You must prepare the data for analysis.
You need to select appropriate methods for producing the experiment.
Which three modules should you run in sequence? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.
NOTE: More than one order of answer choices is correct. You will receive credit for any of the correct orders you select.



Answer:

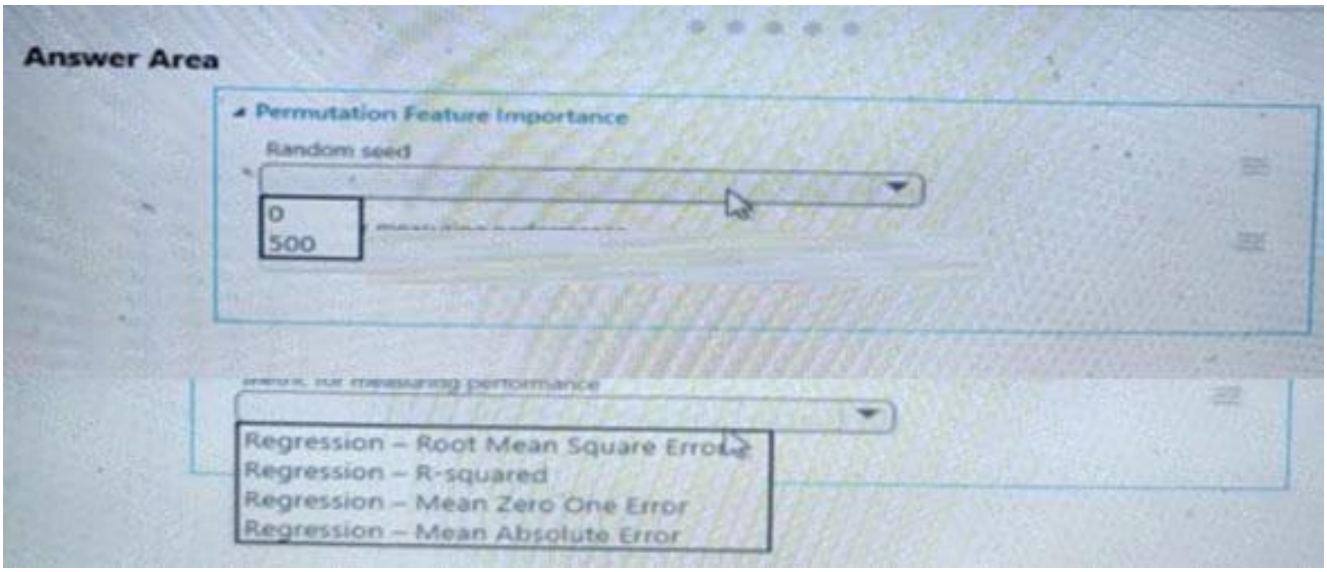


Explanation



The Clean Missing Data module in Azure Machine Learning Studio, to remove, replace, or infer missing values.

NO.90 You need to configure the Permutation Feature Importance module for the model framing requirements.
What should you do? To answer, select the appropriate options in the dialog box in the answer area-
NOTE:
Each correct selection is worth one point.



Answer:

NO.91 You are creating a new experiment in Azure Machine Learning Studio. You have a small dataset that has missing values in many columns. The data does not require the application of predictors for each column. You plan to use the Clean Missing Data module to handle the missing data.
You need to select a data cleaning method.
Which method should you use?

A. Synthetic Minority Oversampling Technique (SMOTE)

B. Replace using MICE

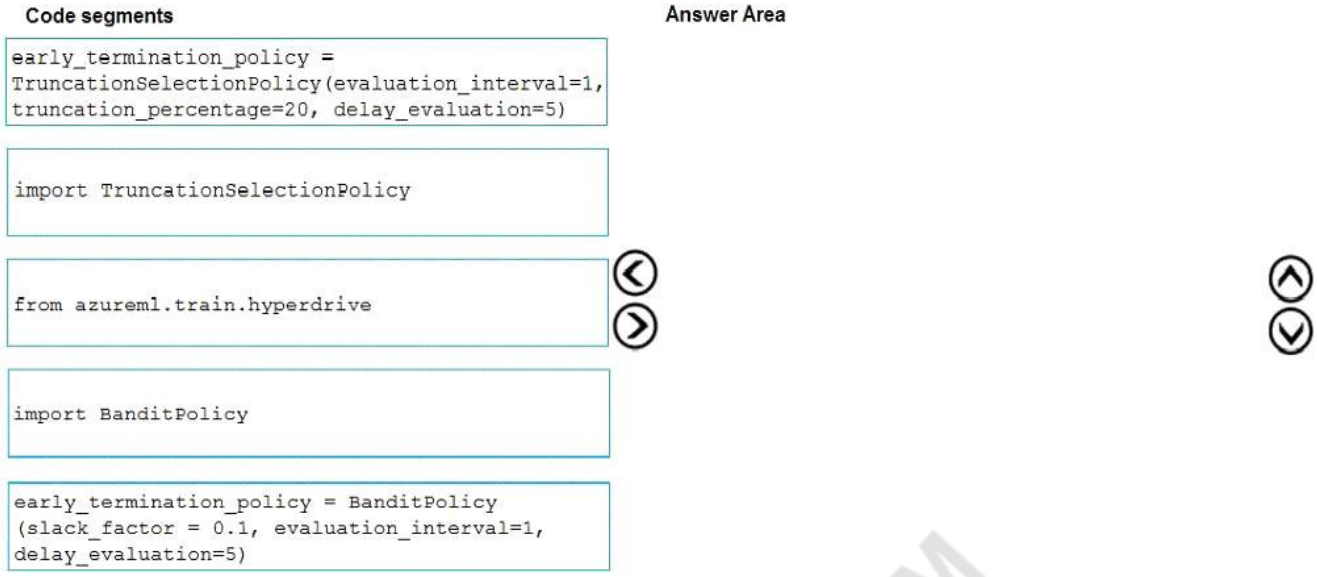
C. Replace using; Probabilistic PCA

D. Normalization

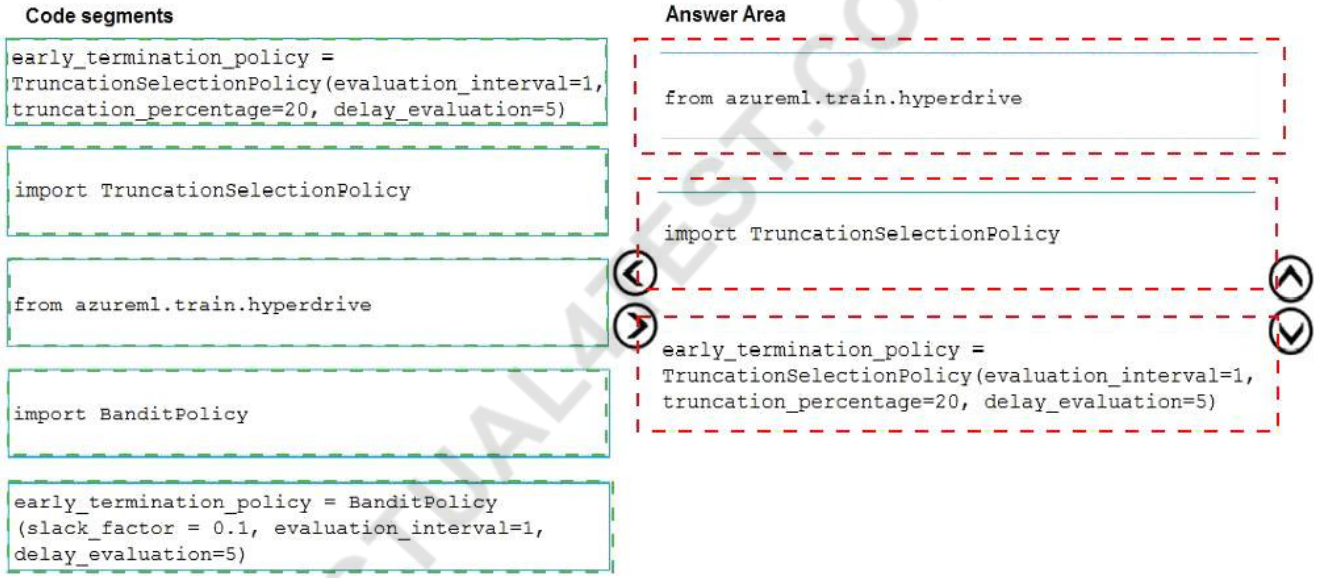
Answer: A

NO.92 You need to implement early stopping criteria as suited in the model training requirements.
Which three code segments should you use to develop the solution? To answer, move the appropriate code segments from the list of code segments to the answer area and arrange them in the correct order.

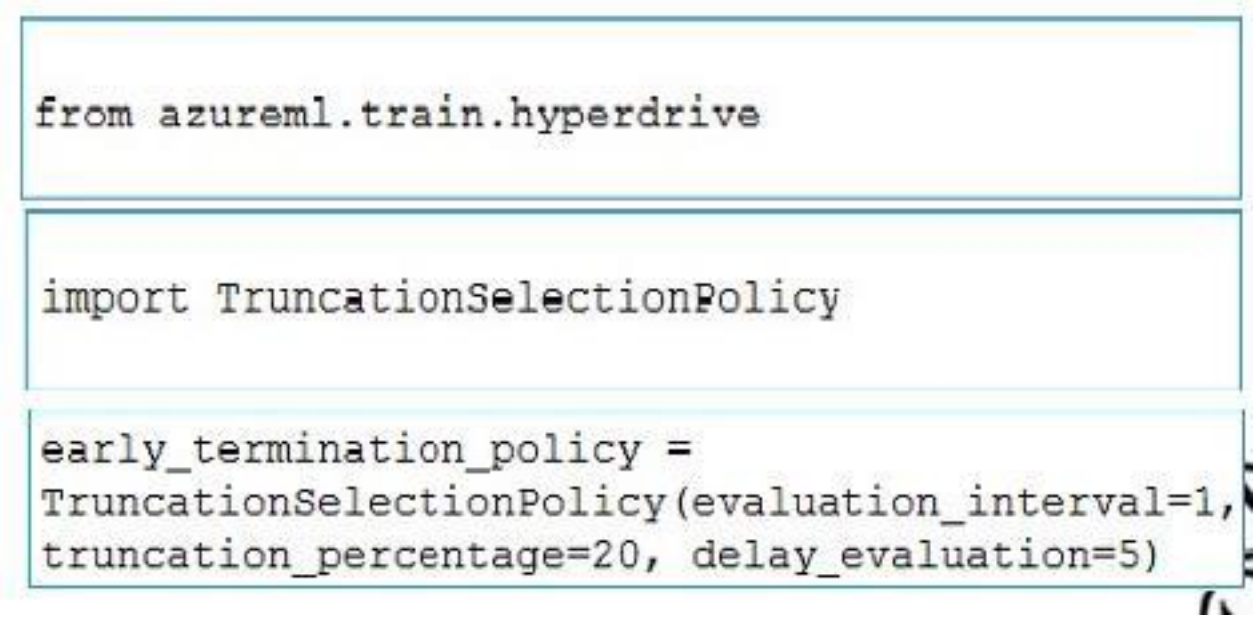
NOTE: More than one order of answer choices is correct. You will receive credit for any of the correct orders you select.



Answer:



Explanation



You need to implement an early stopping criterion on models that provides savings without terminating promising jobs.
Truncation selection cancels a given percentage of lowest performing runs at each evaluation interval. Runs are compared based on their performance on the primary metric and the lowest X% are terminated.
Example:
from azureml.train.hyperdrive import TruncationSelectionPolicy
early_termination_policy = TruncationSelectionPolicy(evaluation_interval=1,
truncation_percentage=20, delay_evaluation=5)

NO.93 You plan to deliver a hands-on workshop to several students. The workshop will focus on creating data visualizations using Python. Each student will use a device that has internet access. Student devices are not configured for Python development. Students do not have administrator access to install software on their devices. Azure subscriptions are not available for students.
You need to ensure that students can run Python-based data visualization code.
Which Azure tool should you use?
A. Anaconda Data Science Platform
B. Azure BatchAI
C. Azure Notebooks
D. Azure Machine Learning Service
Answer: C
Explanation
References:
https://notebooks.azure.com/

NO.94 You need to select a feature extraction method.
Which method should you use?
A. Mutual information
B. Mood's median test
C. Kendall correlation

D. Permutation Feature Importance

Answer: C

Explanation

In statistics, the Kendall rank correlation coefficient, commonly referred to as Kendall's tau coefficient (after the Greek letter), is a statistic used to measure the ordinal association between two measured quantities.

It is a supported method of the Azure Machine Learning Feature selection.

Scenario: When you train a Linear Regression module using a property dataset that shows data for property prices for a large city, you need to determine the best features to use in a model. You can choose standard metrics provided to measure performance before and after the feature importance process completes. You must ensure that the distribution of the features across multiple training models is consistent.

References:

https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/feature-selection-modules

NO.95 You are creating a binary classification by using a two-class logistic regression model.

You need to evaluate the model results for imbalance.

Which evaluation metric should you use?

A. Relative Absolute Error

B. AUC Curve

C. Mean Absolute Error

D. Relative Squared Error

Answer: B

Explanation

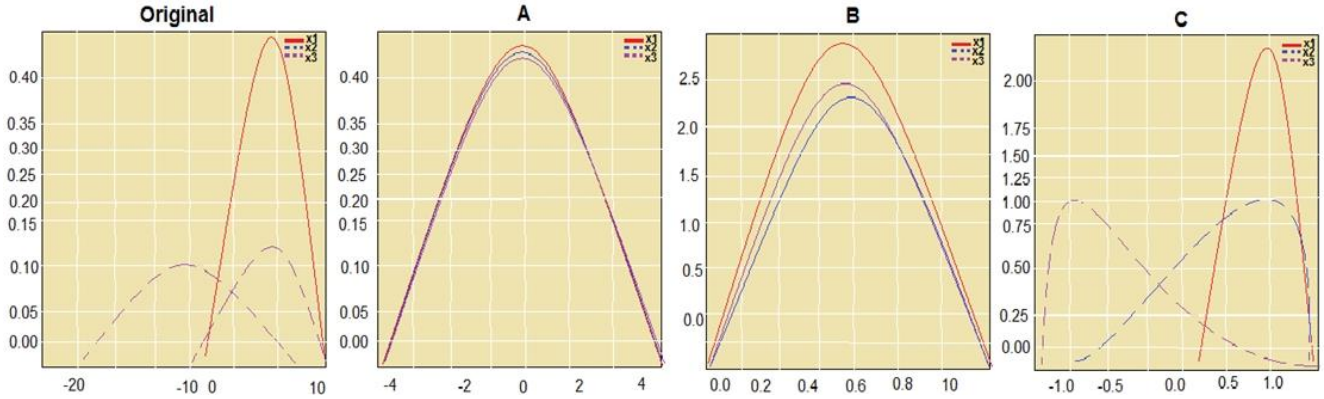
One can inspect the true positive rate vs. the false positive rate in the Receiver Operating Characteristic (ROC) curve and the corresponding Area Under the Curve (AUC) value. The closer this curve is to the upper left corner, the better the classifier's performance is (that is maximizing the true positive rate while minimizing the false positive rate). Curves that are close to the diagonal of the plot, result from classifiers that tend to make predictions that are close to random guessing.

References:

https://docs.microsoft.com/en-us/azure/machine-learning/studio/evaluate-model-performance#evaluating-a-binar

NO.96 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.



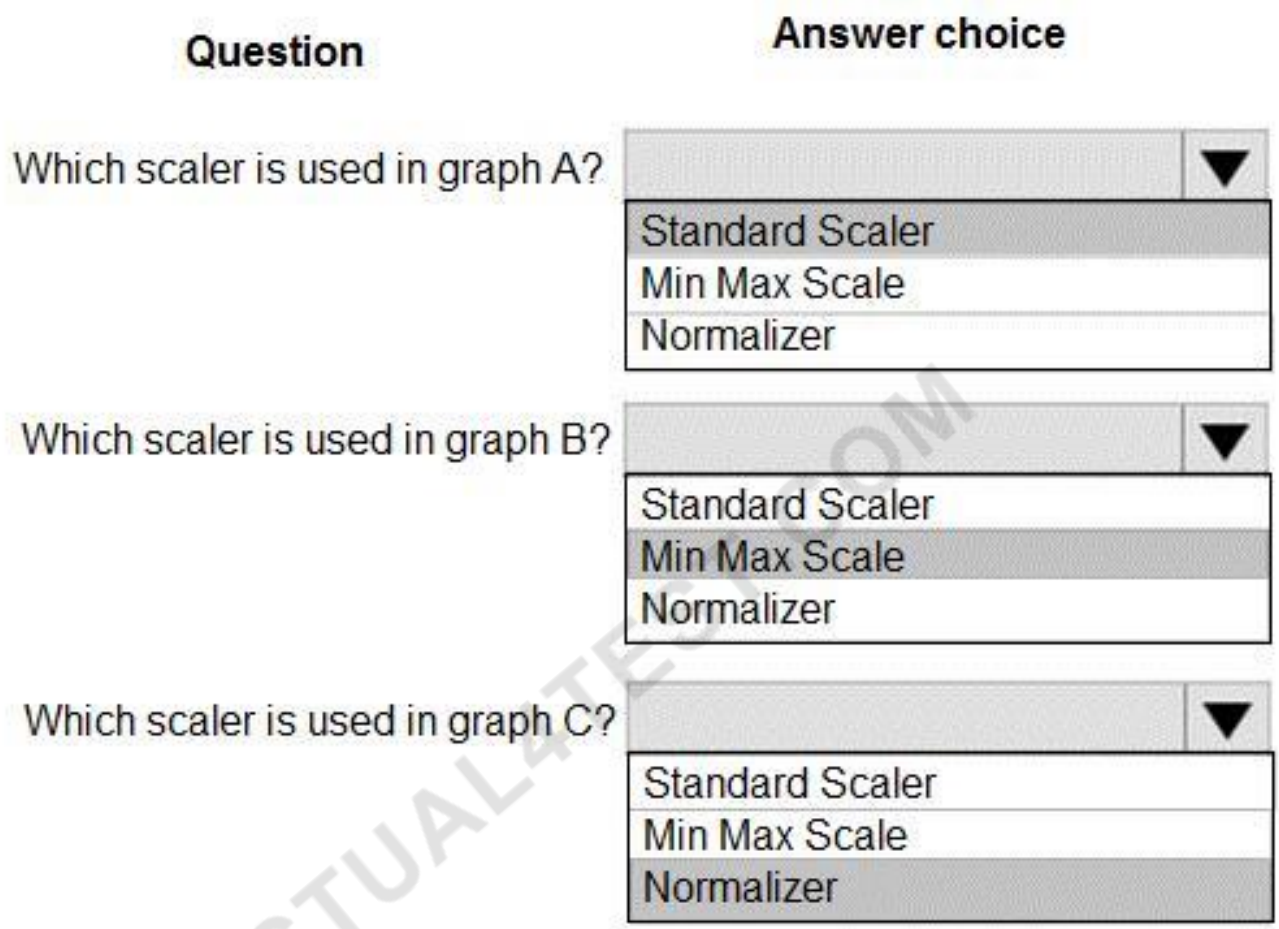
Use the drop-down menus to select the answer choice that answers each question based on the information presented in the graphic.
NOTE: Each correct selection is worth one point.

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:

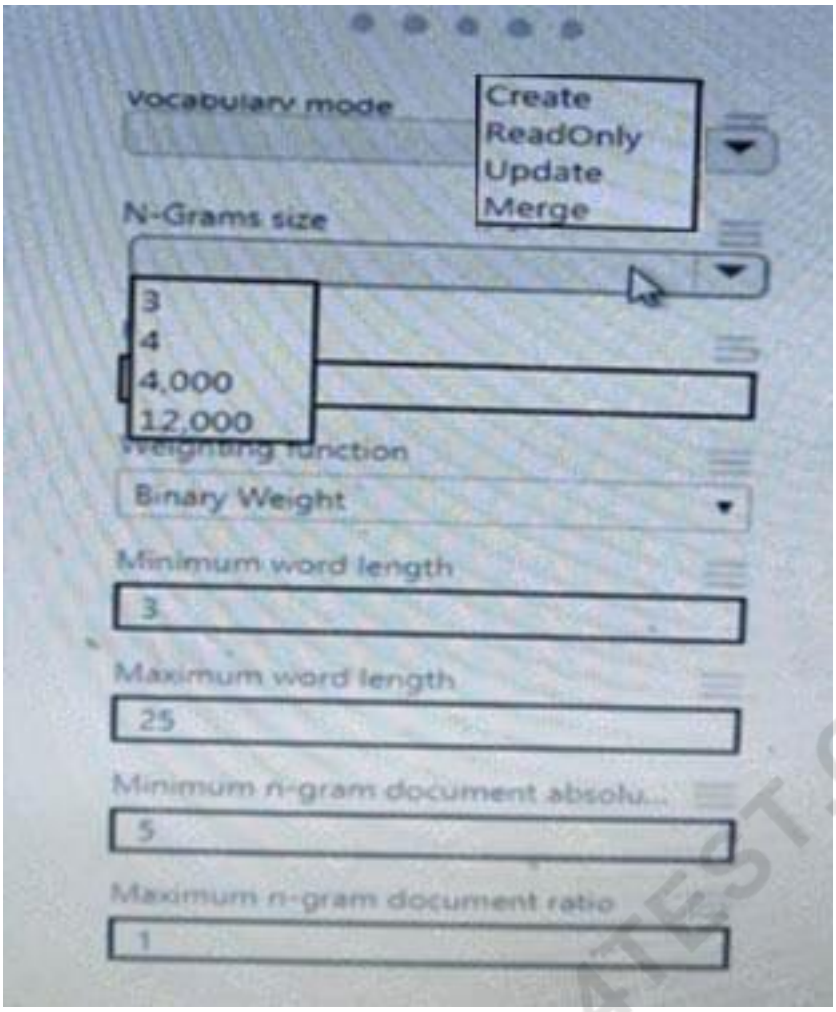
Question	Answer choice
Which scaler is used in graph A?	Standard Scaler Min Max Scale Normalizer
Which scaler is used in graph B?	Standard Scaler Min Max Scale Normalizer
Which scaler is used in graph C?	Standard Scaler Min Max Scale Normalizer

Explanation

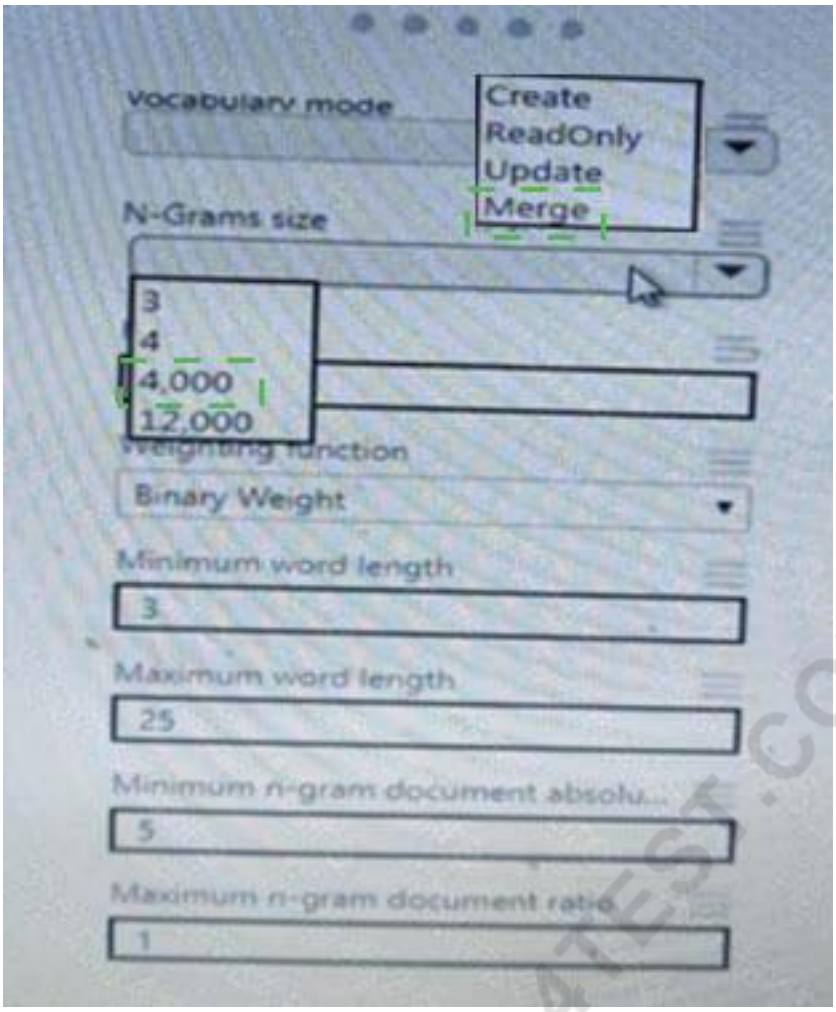


Box 1: StandardScaler
The StandardScaler assumes your data is normally distributed within each feature and will scale them such that the distribution is now centred around 0, with a standard deviation of 1.
Example:
All features are now on the same scale relative to one another.
Box 2: Min Max Scaler
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/

NO.97 You are performing sentiment analysis using a CSV file that includes 12.000 customer reviews written in a short sentence format. You add the CSV file to Azure Machine Learning Studio and Configure it as the starting point dataset of an experiment. You add the Extract N-Gram Features from Text module to the experiment to extract key phrases from the customer review column in the dataset.
You must create a new n-gram text dictionary from the customer review text and set the maximum n-gram size to trigrams.
You need to configure the Extract N Gram features from Text module.
What should you select? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.



Answer:



NO.98 You are solving a classification task.
The dataset is imbalanced.
You need to select an Azure Machine Learning Studio module to improve the classification accuracy.
Which module should you use?
A. Fisher Linear Discriminant Analysis.
B. Filter Based Feature Selection
C. Synthetic Minority Oversampling Teachnique (SMOTE)
D. Permutation Feature Importance
Answer: A

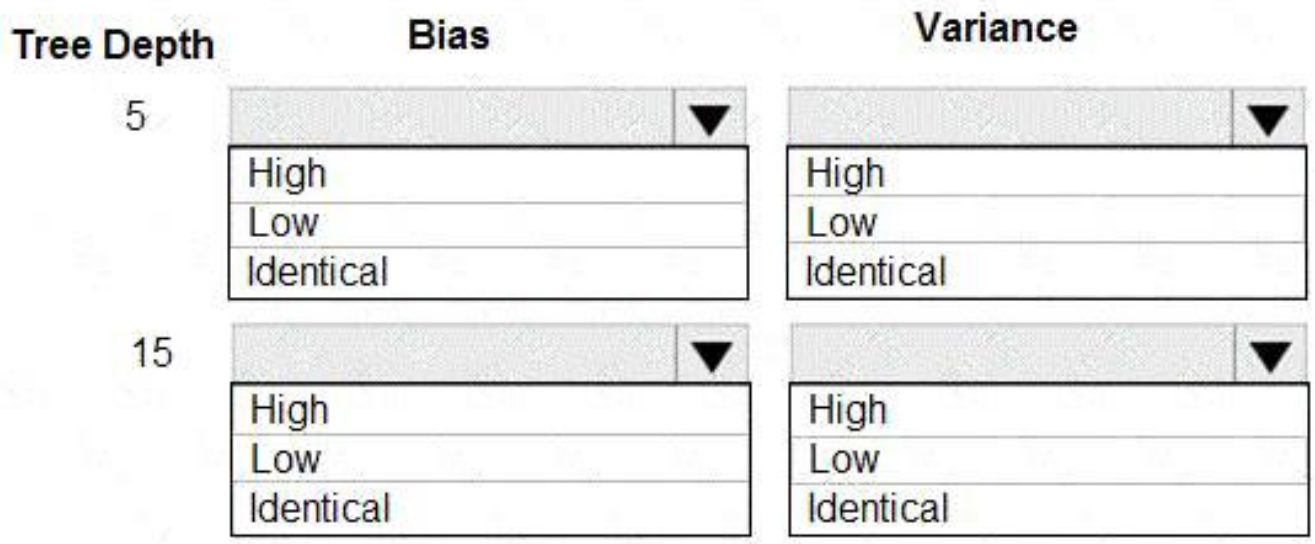
NO.99 You plan to build a team data science environment. Data for training models in machine learning pipelines will be over 20 GB in size.
You have the following requirements:
* Models must be built using Caffe2 or Chainer frameworks.
* Data scientists must be able to use a data science environment to build the machine learning pipelines and train models on their personal devices in both connected and disconnected network environments.
* Personal devices must support updating machine learning pipelines when connected to a network.
You need to select a data science environment.
Which environment should you use?

A. Azure Machine Learning Service
B. Azure Machine Learning Studio
C. Azure Databricks
D. Azure Kubernetes Service (AKS)
Answer: A
Explanation
The Data Science Virtual Machine (DSVM) is a customized VM image on Microsoft's Azure cloud built specifically for doing data science. Caffe2 and Chainer are supported by DSVM.
DSVM integrates with Azure Machine Learning.

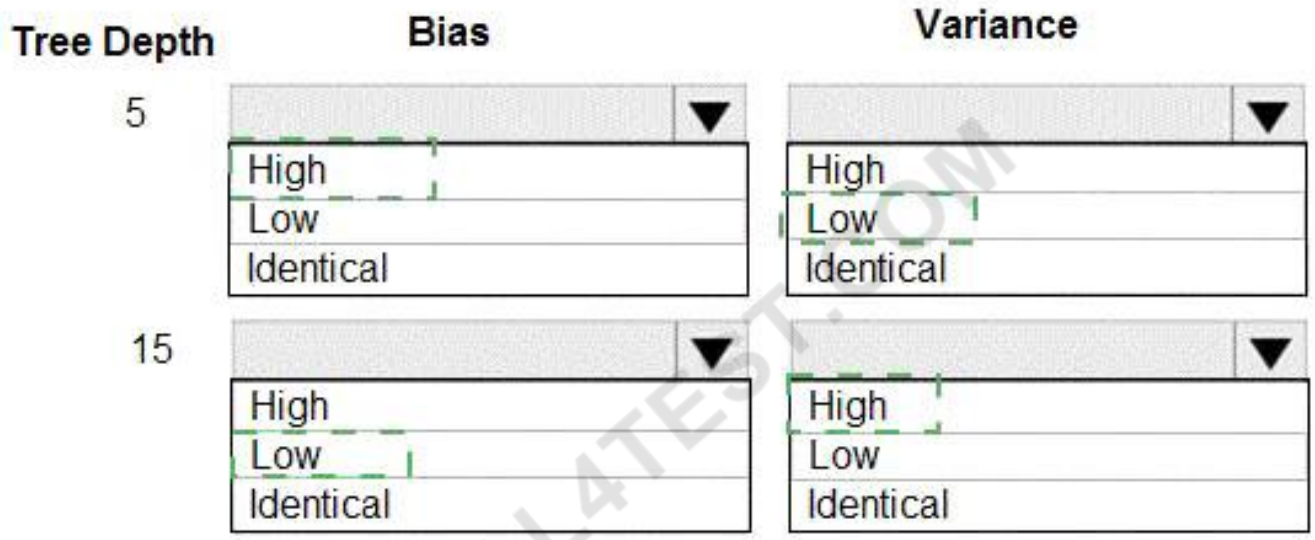
NO.100 You need to select a pre built development environment for a series of data science experiments. You must use the R language for the experiments.
Which three environments can you use? Each correct answer presents a complete solution. NOTE: Each correct selection is worth one point.
A. ML.NET Library on a local environment
B. Azure Machine Learning Studio
C. Data Science Virtual Machine (OSVM)
D. Azure Data bricks
E. Azure Cognitive Services
Answer: A B D

NO.101 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 creating a new experiment in Azure Machine Learning Studio.
One class has a much smaller number of observations than tin- other classes in the training set.
You need to select an appropriate data sampling strategy to compensate for the class imbalance.
Solution: You use the Principal Components Analysis (PCA) sampling mode.
Does the solution meet the goal?
A. Yes
B. No
Answer: B

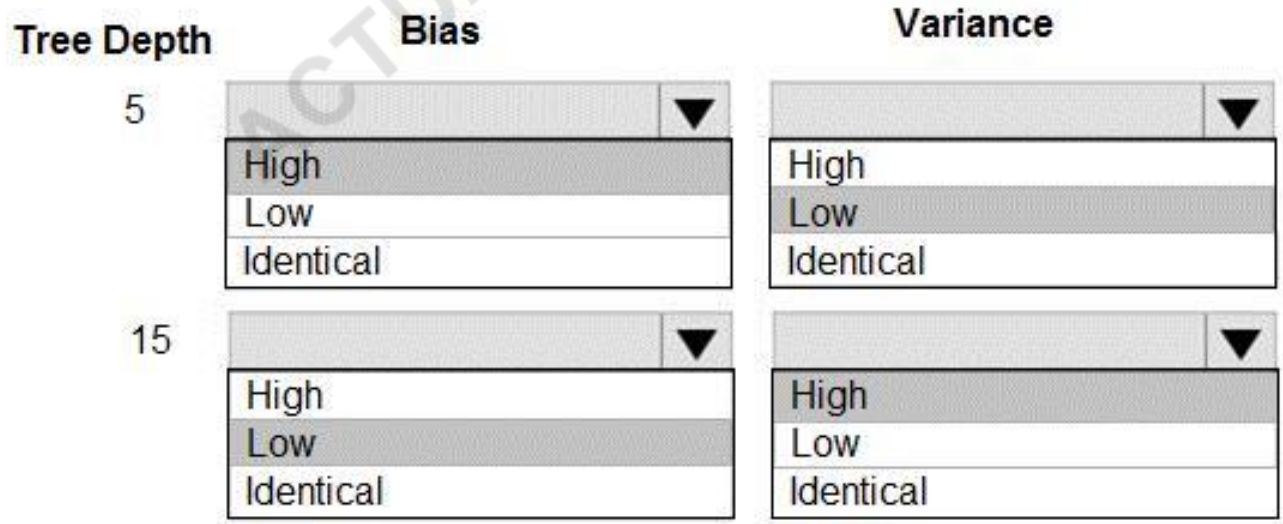
NO.102 You are using a decision tree algorithm. You have trained a model that generalizes well at a tree depth equal to
10.
You need to select the bias and variance properties of the model with varying tree depth values.
Which properties should you select for each tree depth? To answer, select the appropriate options in the answer area.



Answer:



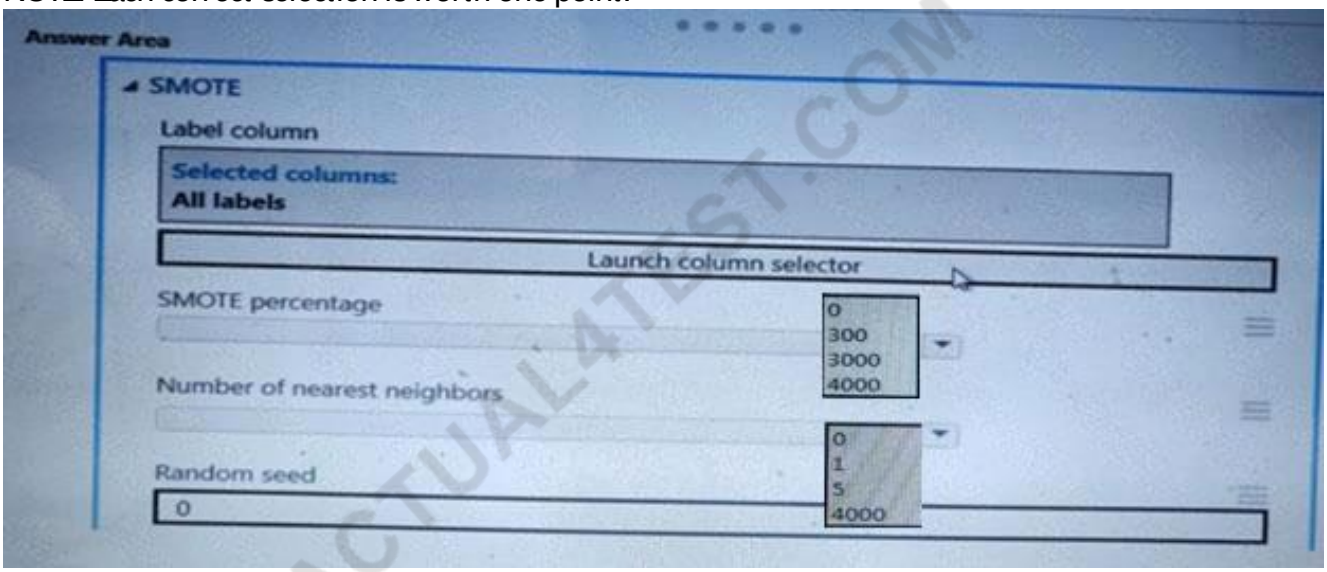
Explanation



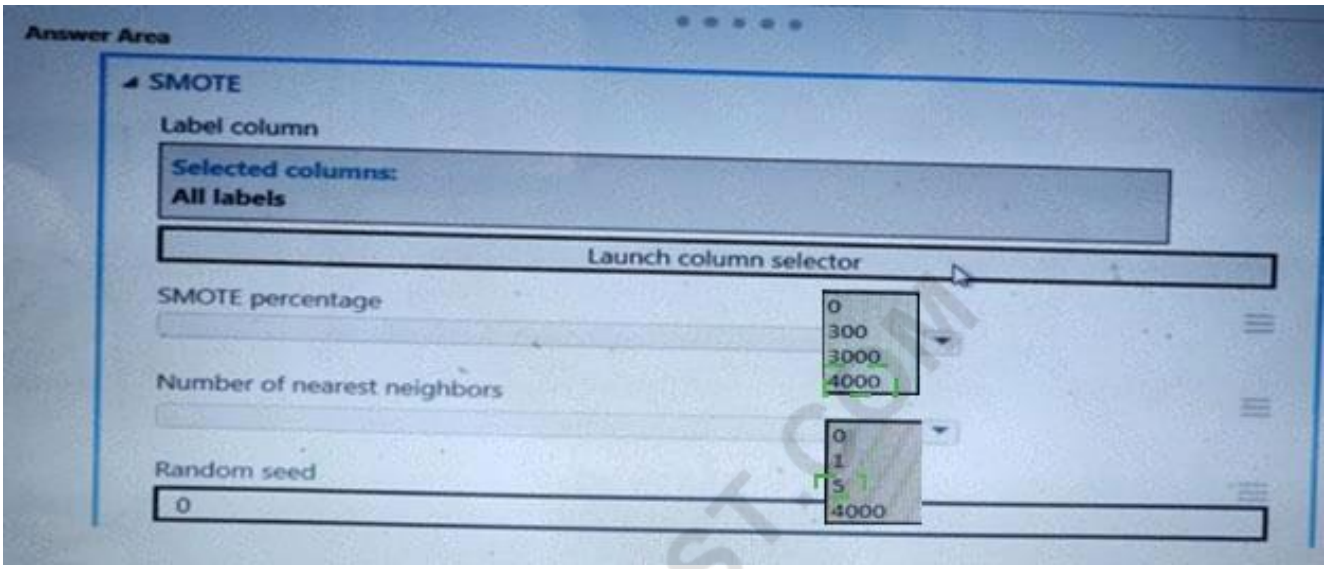
In decision trees, the depth of the tree determines the variance. A complicated decision tree (e.g.

deep) has low bias and high variance.
Note: In statistics and machine learning, the bias-variance 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/

NO.103 You create an experiment in Azure Machine Learning Studio- You add a training dataset that contains 10.000 rows. The first 9.000 rows represent class 0 (90 percent). The first 1.000 rows represent class 1 (10 percent).
The training set is unbalanced between two Classes. You must increase the number of training examples for class 1 to 4,000 by using data rows. You add the Synthetic Minority Oversampling Technique (SMOTE) module to the experiment.
You need to configure the module.
Which values should you use? To answer, select the appropriate options in the dialog box in the answer area.
NOTE: Each correct selection is worth one point.



Answer:



NO.104 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 a data scientist using Azure Machine Learning Studio.
You need to normalize values to produce an output column into bins to predict a target column.
Solution: Apply an Equal Width with Custom Start and Stop binning mode.
Does the solution meet the goal?

A. Yes

B. No

Answer: B

Explanation
Use the Entropy MDL binning mode which has a target column.
References:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/group-data-into-bins

NO.105 You are moving a large dataset from Azure Machine Learning Studio to a Weka environment.
You need to format the data for the Weka environment.
Which module should you use?

A. Convert to CSV

B. Convert to Dataset

C. Convert to ARFF

D. Convert to SVMLight

Answer: C

Explanation
Use the Convert to ARFF module in Azure Machine Learning Studio, to convert datasets and results in Azure Machine Learning to the attribute-relation file format used by the Weka toolset. This format is known as ARFF.

The ARFF data specification for Weka supports multiple machine learning tasks, including data preprocessing, classification, and feature selection. In this format, data is organized by entites and their attributes, and is contained in a single text file.
References:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/convert-to-arff

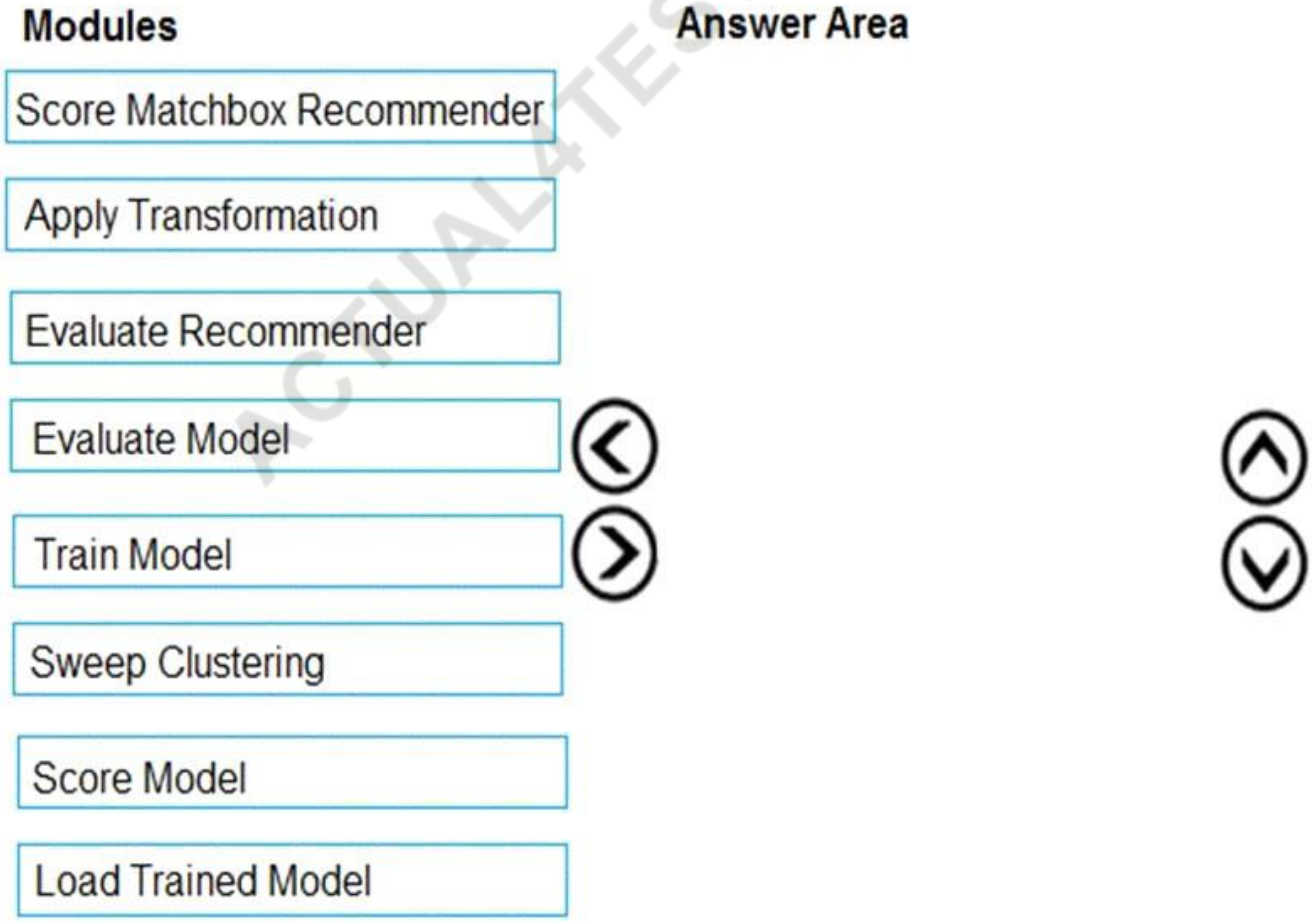
NO.106 You plan to use a Data Science Virtual Machine (DSVM) with the open source deep learning frameworks Caffe2 and Theano. You need to select a pre configured DSVM to support the framework.
What should you create?
A. Data Science Virtual Machine for Linux (CentOS)
B. Data Science Virtual Machine for Windows 2012
C. Data Science Virtual Machine for Windows 2016
D. Geo AI Data Science Virtual Machine with ArcGIS
E. Data Science Virtual Machine for Linux (Ubuntu)
Answer: A

NO.107 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 creating a new experiment in Azure Machine Learning Studio.
One class has a much smaller number of observations than the other classes in the training set.
You need to select an appropriate data sampling strategy to compensate for the class imbalance.
Solution: You use the Scale and Reduce sampling mode.
Does the solution meet the goal?
A. Yes
B. No
Answer: B

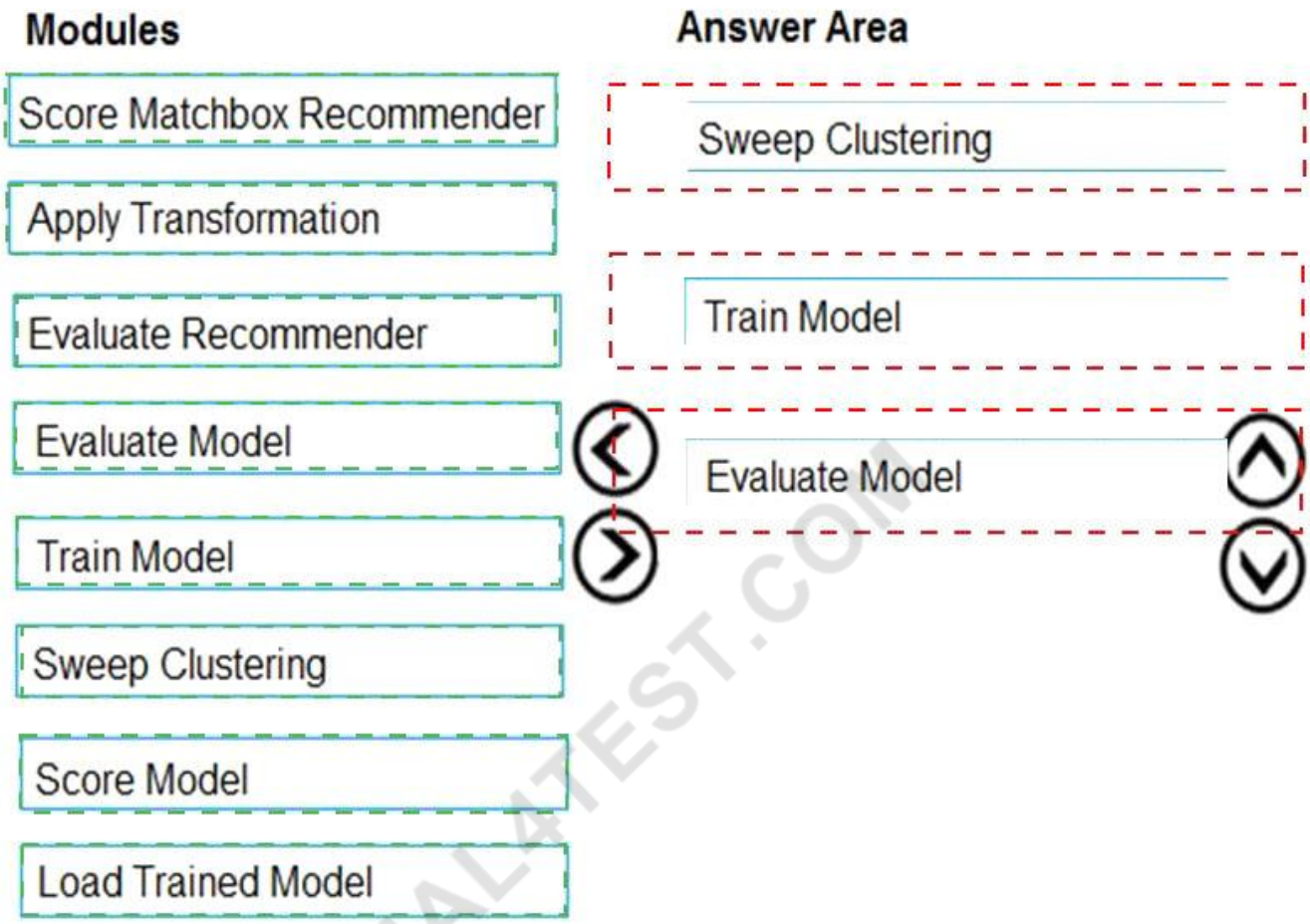
NO.108 You must store data in Azure Blob Storage to support Azure Machine Learning.
You need to transfer the data into Azure Blob Storage.
What are three possible ways to achieve the goal? Each correct answer presents a complete solution.
NOTE: Each correct selection is worth one point.
A. Bulk Insert SQL Query
B. AzCopy
C. Python script
D. Azure Storage Explorer
E. Bulk Copy Program (BCP)
Answer: B C D
Explanation
You can move data to and from Azure Blob storage using different technologies:
Azure Storage-Explorer
AzCopy

Python
SSIS
References:
https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/move-azure-blob

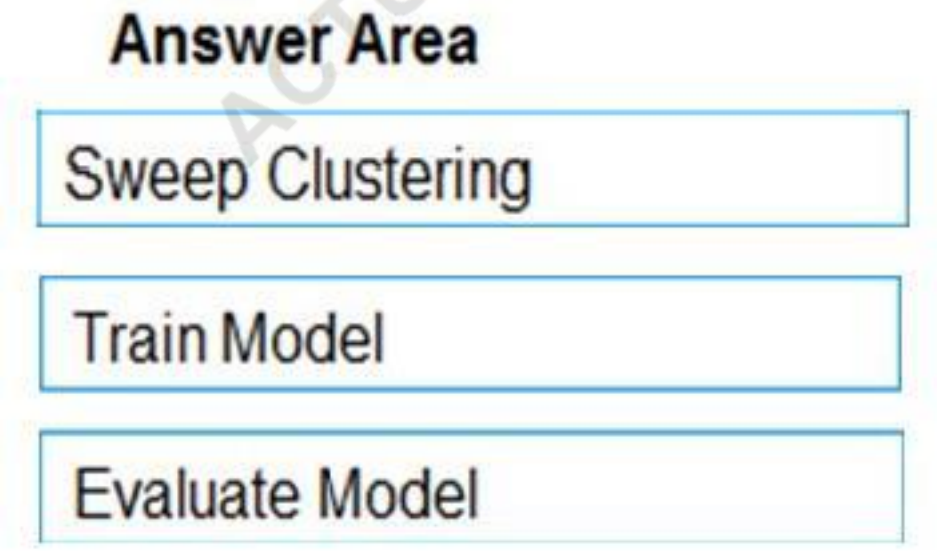
NO.109 You need to produce a visualization for the diagnostic test evaluation according to the data visualization requirements.
Which three modules should you recommend be used in sequence? To answer, move the appropriate modules from the list of modules to the answer area and arrange them in the correct order.



Answer:



Explanation



Step 1: Sweep Clustering
Start by using the "Tune Model Hyperparameters" module to select the best sets of parameters for each of the models we're considering.
One of the interesting things about the "Tune Model Hyperparameters" module is that it not only outputs the results from the Tuning, it also outputs the Trained Model.
Step 2: Train Model
Step 3: Evaluate Model

Scenario: You need to provide the test results to the Fabrikam Residences team. You create data visualizations to aid in presenting the results.
You must produce a Receiver Operating Characteristic (ROC) curve to conduct a diagnostic test evaluation of the model. You need to select appropriate methods for producing the ROC curve in Azure Machine Learning Studio to compare the Two-Class Decision Forest and the Two-Class Decision Jungle modules with one another.
References:
http://breaking-bi.blogspot.com/2017/01/azure-machine-learning-model-evaluation.html

NO.110 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.

Question	Answer choice
Which algorithm minimizes differences between actual and predicted values?	Bayesian Linear Regression Neural Network Regression Boosted Decision Tree Regression Linear Regression Decision Forest Regression
Which approach should you use to find the best parameters for a Linear Regression model for the Online Gradient Descent method?	Set the Decrease learning rate option to True. Set the Decrease learning rate option to True. Set the Create trainer mode option to Parameter Range. Increase the number of epochs. Decrease the number of epochs.

Answer:

Question	Answer choice
Which algorithm minimizes differences between actual and predicted values?	Bayesian Linear Regression Neural Network Regression Boosted Decision Tree Regression Linear Regression Decision Forest Regression
Which approach should you use to find the best parameters for a Linear Regression model for the Online Gradient Descent method?	Set the Decrease learning rate option to True. Set the Decrease learning rate option to True. Set the Create trainer mode option to Parameter Range. Increase the number of epochs. Decrease the number of epochs.

Explanation

Question	Answer choice
Which algorithm minimizes differences between actual and predicted values?	Bayesian Linear Regression Neural Network Regression **Boosted Decision Tree Regression** Linear Regression Decision Forest Regression
Which approach should you use to find the best parameters for a Linear Regression model for the Online Gradient Descent method?	Set the Decrease learning rate option to True. Set the Decrease learning rate option to True. **Set the Create trainer mode option to Parameter Range.** Increase the number of epochs. Decrease the number of epochs.

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