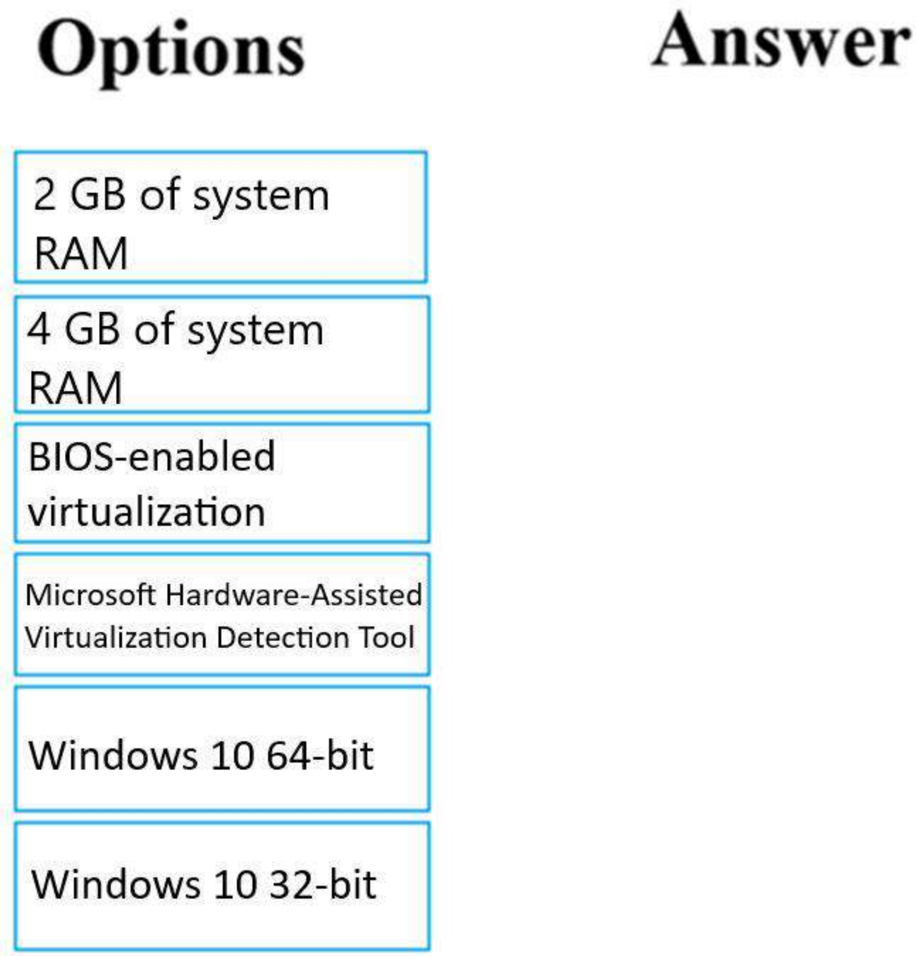
CLOUD CERTIFIED

PREPARE WITH BEST PRACTICE STUDY MATERIAL TO GET SUCCESS

DP-100 Designing and Implementing a Data Science

Question #1*Topic 1*
DRAG DROP -
You are planning to host practical training to acquaint staff with Docker for Windows.
Staff devices must support the installation of Docker.
Which of the following are requirements for this installation? Answer by dragging the correct options from the list to the answer area.

Select and Place:



Correct Answer:

Options	Answer
2 GB of system RAM	4 GB of system RAM
	BIOS-enabled virtualization
	Windows 10 64-bit
Microsoft Hardware-Assisted Virtualization Detection Tool	
Windows 10 32-bit	

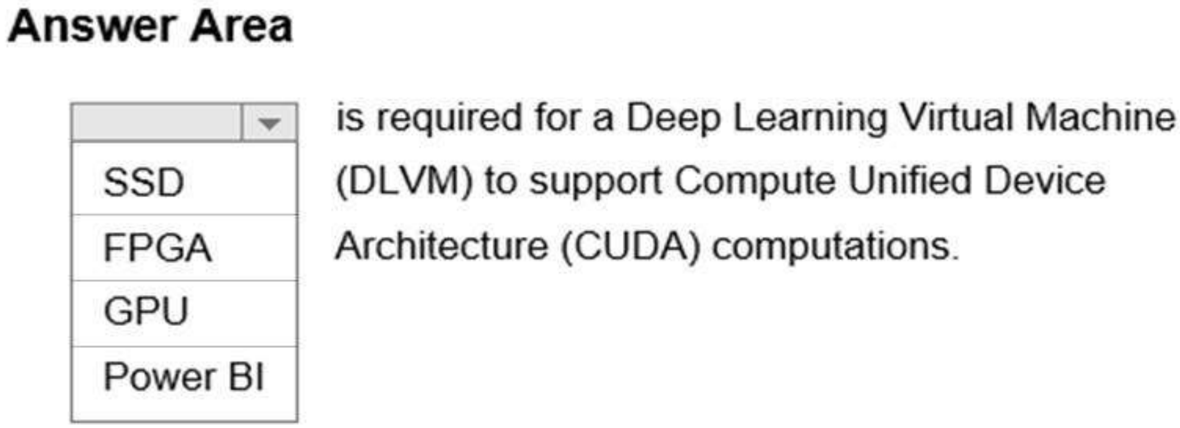
Reference:
https://docs.docker.com/toolbox/toolbox_install_windows/
https://blogs.technet.microsoft.com/canitpro/2015/09/08/step-by-step-enabling-hyper-v-for-use-on-windows-10/ https://docs.docker.com/docker-for-windows/install/

WSL 2 backend

- Windows 10 64-bit: Home or Pro 2004 (build 19041) or higher, or Enterprise or Education 1909 (build 18363) or higher.
- Enable the WSL 2 feature on Windows. For detailed instructions, refer to the Microsoft documentation.
- The following hardware prerequisites are required to successfully run WSL 2 on Windows 10:
  - 64-bit processor with Second Level Address Translation (SLAT)
  - 4GB system RAM
  - BIOS-level hardware virtualization support must be enabled in the BIOS settings. For more information, see Virtualization.
- Download and install the Linux kernel update package.

Question #2*Topic 1*

HOTSPOT -
Complete the sentence by selecting the correct option in the answer area.
Hot Area:



Correct Answer:

A Deep Learning Virtual Machine is a pre-configured environment for deep learning using GPU instances.

Question #3*Topic 1*
You need to implement a Data Science Virtual Machine (DSVM) that supports the Caffe2 deep learning framework.
Which of the following DSVM should you create?

- A. Windows Server 2012 DSVM
- B. Windows Server 2016 DSVM
- C. Ubuntu 16.04 DSVM
- D. CentOS 7.4 DSVM

Correct Answer: *C*
Caffe2 is supported by Data Science Virtual Machine for Linux.
Microsoft offers Linux editions of the DSVM on Ubuntu 16.04 LTS and CentOS 7.4.
However, only the DSVM on Ubuntu is preconfigured for Caffe2.
Reference:

https://docs.microsoft.com/en-us/azure/machine-learning/data-science-virtual-machine/overview

Comparison with Azure Machine Learning

The DSVM is a customized VM image for Data Science but Azure Machine Learning (AzureML) is an end-to-end platform that encompasses:

- Fully Managed Compute
  - Compute Instances
  - Compute Clusters for distributed ML tasks
  - Inference Clusters for real-time scoring
- Datastores (for example Blob, ADLS Gen2, SQL DB)
- Experiment tracking
- Model management
- Notebooks
- Environments (manage conda and R dependencies)
- Labeling
- Pipelines (automate End-to-End Data science workflows)

Question #4*Topic 1*
This question is included in a number of questions that depicts the identical set-up. However, every question has a distinctive result. Establish if the recommendation satisfies the requirements.
You have been tasked with employing a machine learning model, which makes use of a PostgreSQL database and needs GPU processing, to forecast prices.
You are preparing to create a virtual machine that has the necessary tools built into it.
You need to make use of the correct virtual machine type.
Recommendation: You make use of a Geo AI Data Science Virtual Machine (Geo-DSVM) Windows edition.
Will the requirements be satisfied?

- A. Yes
- B. No

Correct Answer: *B*
The Azure Geo AI Data Science VM (Geo-DSVM) delivers geospatial analytics capabilities from Microsoft's Data Science VM. Specifically, this VM extends the AI and data science toolkits in the Data Science VM by adding ESRI's market-leading ArcGIS Pro Geographic Information System.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/data-science-virtual-machine/overview

Question #5*Topic 1*
This question is included in a number of questions that depicts the identical set-up. However, every question has a distinctive result. Establish if the recommendation satisfies the requirements.
You have been tasked with employing a machine learning model, which makes use of a PostgreSQL database and needs GPU processing, to forecast prices.
You are preparing to create a virtual machine that has the necessary tools built into it.
You need to make use of the correct virtual machine type.

Recommendation: You make use of a Deep Learning Virtual Machine (DLVM) Windows edition.
Will the requirements be satisfied?

- A. Yes
- B. No

Correct Answer: *A*
DLVM is a template on top of DSVM image. In terms of the packages, GPU drivers etc are all there in the DSVM image. Mostly it is for convenience during creation where we only allow DLVM to be created on GPU VM instances on Azure.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/data-science-virtual-machine/overview

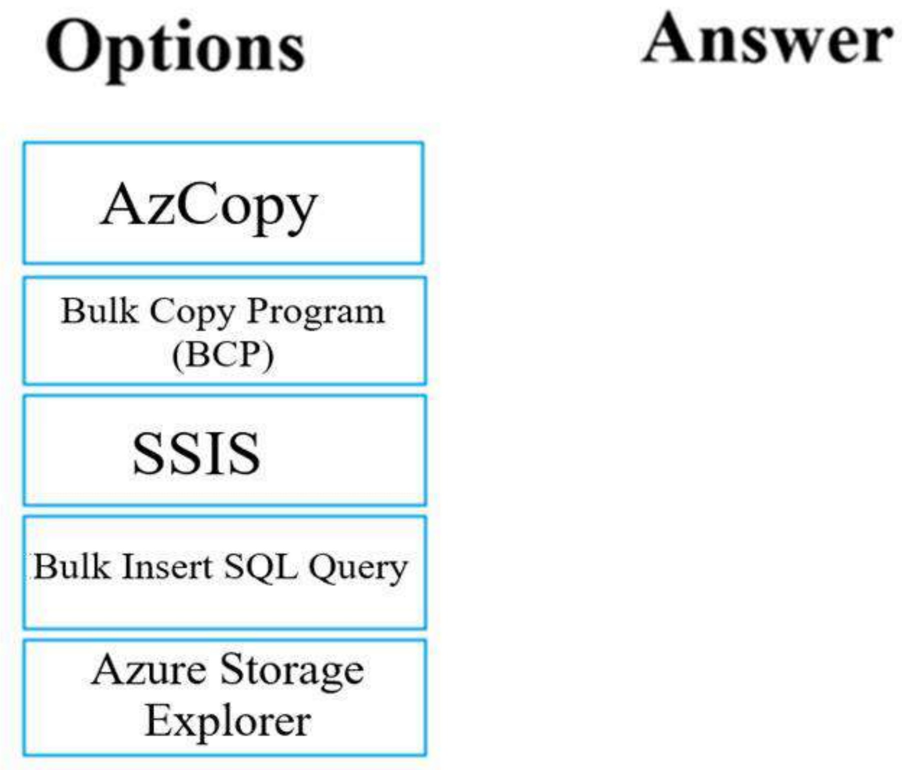
Question #6*Topic 1*
This question is included in a number of questions that depicts the identical set-up. However, every question has a distinctive result. Establish if the recommendation satisfies the requirements.
You have been tasked with employing a machine learning model, which makes use of a PostgreSQL database and needs GPU processing, to forecast prices.
You are preparing to create a virtual machine that has the necessary tools built into it.
You need to make use of the correct virtual machine type.
Recommendation: You make use of a Data Science Virtual Machine (DSVM) Windows edition.
Will the requirements be satisfied?

- A. Yes
- B. No

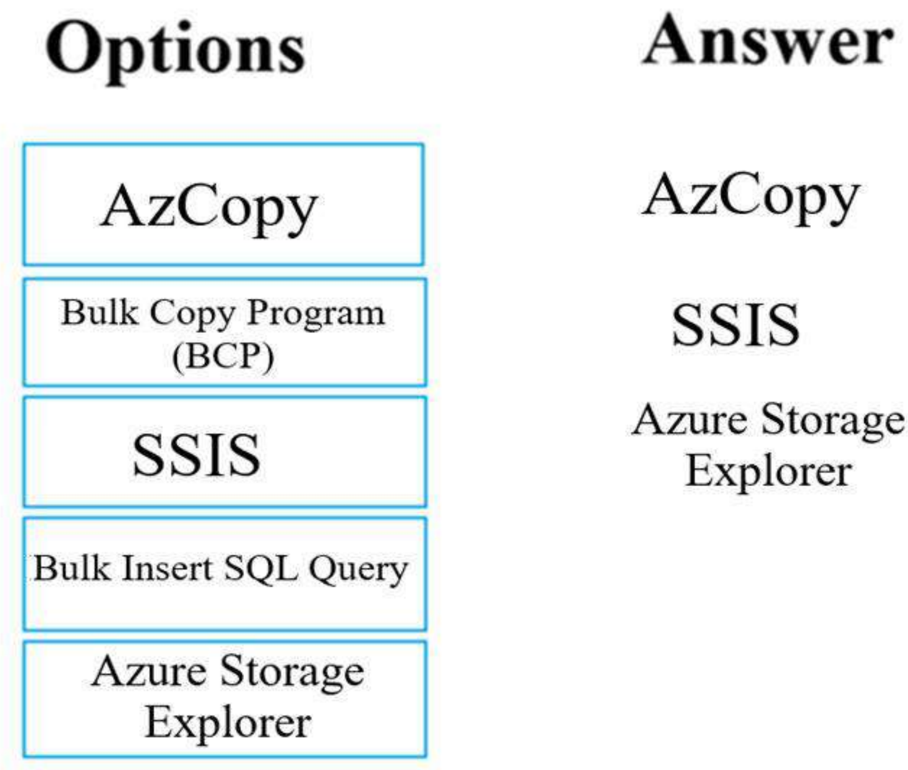
Correct Answer: *A*
In the DSVM, your training models can use deep learning algorithms on hardware that's based on graphics processing units (GPUs).
PostgreSQL is available for the following operating systems: Linux (all recent distributions), 64-bit installers available for macOS (OS X) version 10.6 and newer ג€"
Windows (with installers available for 64-bit version; tested on latest versions and back to Windows 2012 R2.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/data-science-virtual-machine/overview

Question #7*Topic 1*
DRAG DROP -
You have been tasked with moving data into Azure Blob Storage for the purpose of supporting Azure Machine Learning.
Which of the following can be used to complete your task? Answer by dragging the correct options from the list to the answer area.

Select and Place:



Correct Answer:



You can move data to and from Azure Blob storage using different technologies:

⊸ Azure Storage-Explorer

⊸ AzCopy

⊸ Python

⊸ SSIS

Reference:

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

Different technologies for moving data

The following articles describe how to move data to and from Azure Blob storage using different technologies.

- Azure Storage-Explorer
- AzCopy
- Python
- SSIS

Which method is best for you depends on your scenario. The Scenarios for advanced analytics in Azure Machine Learning article helps you determine the resources you need for a variety of data science workflows used in the advanced analytics process.

Question #8*Topic 1*

HOTSPOT -

Complete the sentence by selecting the correct option in the answer area.

Hot Area:

Answer Area

To move a large dataset from Azure Machine Learning Studio to a Weka environment, the data must be converted to the [▼] format.

- CSV
- DOCX
- ARFF
- TXT

Correct Answer:

Answer Area

To move a large dataset from Azure Machine Learning Studio to a Weka environment, the data must be converted to the [dropdown] format.

- CSV
- DOCX
- ARFF
- TXT

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 entities and their attributes, and is contained in a single text file.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/convert-to-arff

Module overview

This article describes how to use the Convert to ARFF module in Machine Learning Studio (classic), to convert datasets and results 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. You can find details of the Weka file format in the Technical Notes section.

In general, conversion to the Weka file format is required only if you want to use both Machine Learning and Weka, and intend to move your training data back and forth between them.

For more information about the Weka toolset, see this Wikipedia article: Weka (machine learning)

Question #9*Topic 1*
You have been tasked with designing a deep learning model, which accommodates the most recent edition of Python, to recognize language.
You have to include a suitable deep learning framework in the Data Science Virtual Machine (DSVM).
Which of the following actions should you take?

- A. You should consider including Rattle.
- B. You should consider including TensorFlow.
- C. You should consider including Theano.
- D. You should consider including Chainer.

Correct Answer: *B*
Reference:
https://www.infoworld.com/article/3278008/what-is-tensorflow-the-machine-learning-library-explained.html

Machine learning is a complex discipline. But implementing machine learning models is far less daunting and difficult than it used to be, thanks to machine learning frameworks—such as Google's TensorFlow—that ease the process of acquiring data, training models, serving predictions, and refining future results.

Created by the Google Brain team, TensorFlow is an open source library for numerical computation and large-scale machine learning. TensorFlow bundles together a slew of machine learning and deep learning (aka neural networking) models and algorithms and makes them useful by way of a common metaphor. It uses Python to provide a convenient front-end API for building applications with the framework, while executing those applications in high-performance C++.

Question #10*Topic 1*
This question is included in a number of questions that depicts the identical set-up. However, every question has a distinctive result. Establish if the recommendation satisfies the requirements.
You have been tasked with evaluating your model on a partial data sample via k-fold cross-validation.
You have already configured a k parameter as the number of splits. You now have to configure the k parameter for the cross-validation with the usual value choice.
Recommendation: You configure the use of the value k=3.
Will the requirements be satisfied?

- A. Yes
- B. No

Correct Answer: *B*

Question #11*Topic 1*
This question is included in a number of questions that depicts the identical set-up. However, every question has a distinctive result. Establish if the recommendation satisfies the requirements.
You have been tasked with evaluating your model on a partial data sample via k-fold cross-validation.

You have already configured a k parameter as the number of splits. You now have to configure the k parameter for the cross-validation with the usual value choice.
Recommendation: You configure the use of the value k=10.
Will the requirements be satisfied?

- A. Yes
- B. No

Correct Answer: *A*
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.

Question #12*Topic 1*
You construct a machine learning experiment via Azure Machine Learning Studio.
You would like to split data into two separate datasets.
Which of the following actions should you take?

- A. You should make use of the Split Data module.
- B. You should make use of the Group Categorical Values module.
- C. You should make use of the Clip Values module.
- D. You should make use of the Group Data into Bins module.

Correct Answer: *A*
The Group Data into Bins module supports multiple options for binning data. You can customize how the bin edges are set and how values are apportioned into the bins.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/group-data-

into-bins

Module overview

This article describes how to use the Group Data into Bins module in Machine Learning Studio (classic), to group numbers or change the distribution of continuous data.

The Group Data into Bins module supports multiple options for binning data. You can customize how the bin edges are set and how values are apportioned into the bins. For example, you can:

- Manually type a series of values to serve as the bin boundaries.
- Calculate entropy scores to determine an information values for each range, to optimize the bins in the predictive model. + Assign values to bins by using *quantiles*, or percentile ranks.
- Control the number of values in each bin can also be controlled.
- Force an even distribution of values into the bins.

Question #13*Topic 1*
You have been tasked with creating a new Azure pipeline via the Machine Learning designer.
You have to makes sure that the pipeline trains a model using data in a comma-separated values (CSV) file that is published on a website. A dataset for the file for this file does not exist.
Data from the CSV file must be ingested into the designer pipeline with the least amount of administrative effort as possible.
Which of the following actions should you take?

- A. You should make use of the Convert to TXT module.
- B. You should add the Copy Data object to the pipeline.
- C. You should add the Import Data object to the pipeline.
- D. You should add the Dataset object to the pipeline.

Correct Answer: *C*
The preferred way to provide data to a pipeline is a Dataset object. The Dataset object points to data that lives in or is accessible from a datastore or at a Web
URL. The Dataset class is abstract, so you will create an instance of either a FileDataset (referring to one or more files) or a TabularDataset that's created by from one or more files with delimited columns of data.
Example:
from azureml.core import Dataset
iris_tabular_dataset = Dataset.Tabular.from_delimited_files([(def_blob_store, 'train-dataset/iris.csv')])
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-create-your-first-pipeline

Question #14*Topic 1*
This question is included in a number of questions that depicts the identical set-up. However, every question has a distinctive result. Establish if the recommendation satisfies the requirements.
You are in the process of creating a machine learning model. Your dataset includes rows with null

and missing values.
You plan to make use of the Clean Missing Data module in Azure Machine Learning Studio to detect and fix the null and missing values in the dataset.
Recommendation: You make use of the Replace with median option.
Will the requirements be satisfied?

- A. Yes
- B. No

Correct Answer: *A*
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/clean-missing-data

Question #15*Topic 1*
This question is included in a number of questions that depicts the identical set-up. However, every question has a distinctive result. Establish if the recommendation satisfies the requirements.
You are in the process of creating a machine learning model. Your dataset includes rows with null and missing values.
You plan to make use of the Clean Missing Data module in Azure Machine Learning Studio to detect and fix the null and missing values in the dataset.
Recommendation: You make use of the Custom substitution value option.
Will the requirements be satisfied?

- A. Yes
- B. No

Correct Answer: *A*
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/clean-missing-data

How to use Clean Missing Data

This module lets you define a cleaning operation. You can also save the cleaning operation so that you can apply it later to new data. See the following links for a description of how to create and save a cleaning process:

- To replace missing values
- To apply a cleaning transformation to new data

Important

The cleaning method that you use for handling missing values can dramatically affect your results. We recommend that you experiment with different methods. Consider both the justification for use of a particular method, and the quality of the results.

Question #16*Topic 1*
This question is included in a number of questions that depicts the identical set-up. However, every question has a distinctive result. Establish if the recommendation satisfies the requirements.
You are in the process of creating a machine learning model. Your dataset includes rows with null and missing values.
You plan to make use of the Clean Missing Data module in Azure Machine Learning Studio to detect and fix the null and missing values in the dataset.
Recommendation: You make use of the Remove entire row option.
Will the requirements be satisfied?

- A. Yes
- B. No

Correct Answer: *A*
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.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/clean-missing-data

Question #17*Topic 1*
You need to consider the underlined segment to establish whether it is accurate.
To transform a categorical feature into a binary indicator, you should make use of the Clean Missing Data module.
Select `No adjustment required` if the underlined segment is accurate. If the underlined segment is inaccurate, select the accurate option.

- A. No adjustment required.
- B. Convert to Indicator Values
- C. Apply SQL Transformation
- D. Group Categorical Values

Correct Answer: *B*
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.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/convert-to-indicator-values

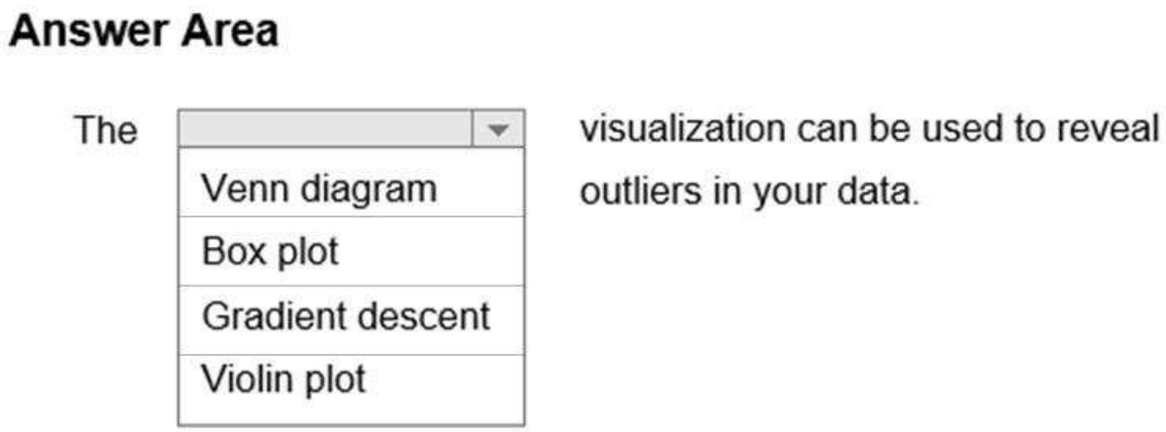
Question #18*Topic 1*
You need to consider the underlined segment to establish whether it is accurate.
To improve the amount of low incidence cases in a dataset, you should make use of the SMOTE module.
Select `No adjustment required` if the underlined segment is accurate. If the underlined segment is inaccurate, select the accurate option.

- A. No adjustment required.
- B. Remove Duplicate Rows

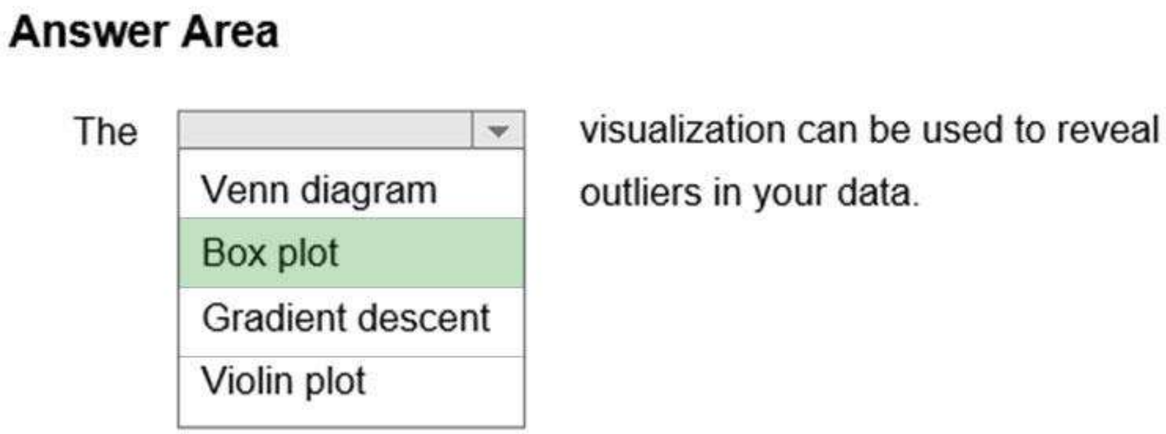
- C. Join Data
- D. Edit Metadata

Correct Answer: *A*
Use the SMOTE module in Azure Machine Learning Studio to increase the number of underrepresented 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.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/smote

Question #19*Topic 1*
HOTSPOT -
You need to consider the underlined segment to establish whether it is accurate.
Hot Area:



Correct Answer:



The box-plot algorithm can be used to display outliers.
Reference:
https://medium.com/analytics-vidhya/what-is-an-outliers-how-to-detect-and-remove-them-which-algorithm-are-sensitive-towards-outliers-2d501993d59

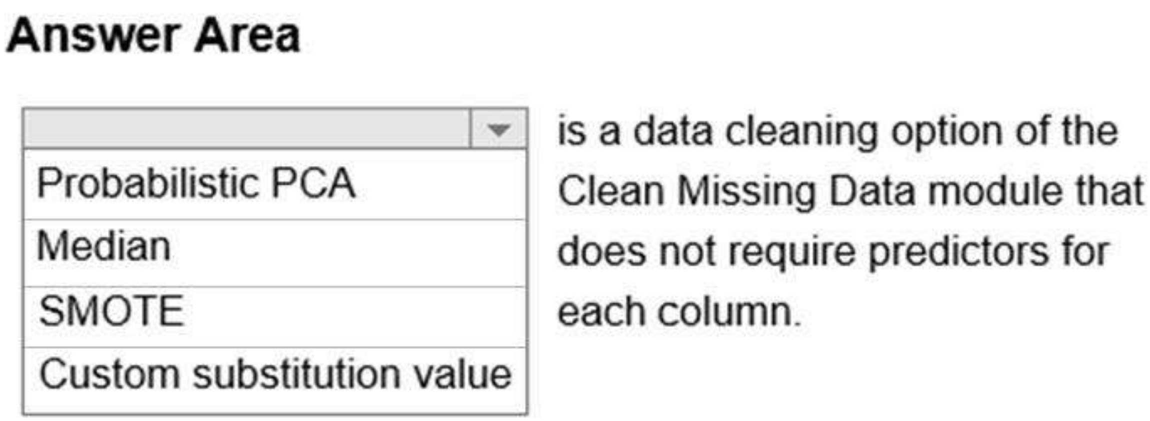
Question #20*Topic 1*

You are planning to host practical training to acquaint learners with data visualization creation using Python. Learner devices are able to connect to the internet.
Learner devices are currently NOT configured for Python development. Also, learners are unable to install software on their devices as they lack administrator permissions. Furthermore, they are unable to access Azure subscriptions.
It is imperative that learners are able to execute Python-based data visualization code.
Which of the following actions should you take?

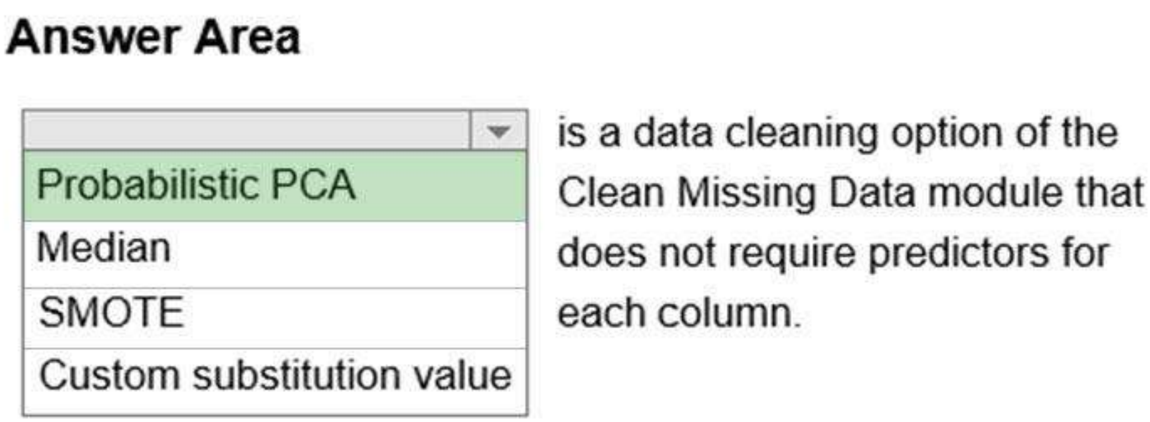
- A. You should consider configuring the use of Azure Container Instance.
- B. You should consider configuring the use of Azure BatchAI.
- C. You should consider configuring the use of Azure Notebooks.
- D. You should consider configuring the use of Azure Kubernetes Service.

Correct Answer: *C*
Reference:
https://notebooks.azure.com/

Question #21*Topic 1*
HOTSPOT -
Complete the sentence by selecting the correct option in the answer area.
Hot Area:



Correct Answer:



Replace using Probabilistic PCA: Compared to other options, such as Multiple Imputation using Chained Equations (MICE), this option has the advantage of not requiring the application of predictors for each column. Instead, it approximates the covariance for the full dataset. Therefore, it might offer better performance for datasets that have missing values in many columns.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/clean-missing-data

Question #22*Topic 1*
You have recently concluded the construction of a binary classification machine learning model. You are currently assessing the model. You want to make use of a visualization that allows for precision to be used as the measurement for the assessment.
Which of the following actions should you take?

- A. You should consider using Venn diagram visualization.
- B. You should consider using Receiver Operating Characteristic (ROC) curve visualization.
- C. You should consider using Box plot visualization.
- D. You should consider using the Binary classification confusion matrix visualization.

Correct Answer: *D*
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-understand-automated-ml#confusion-matrix

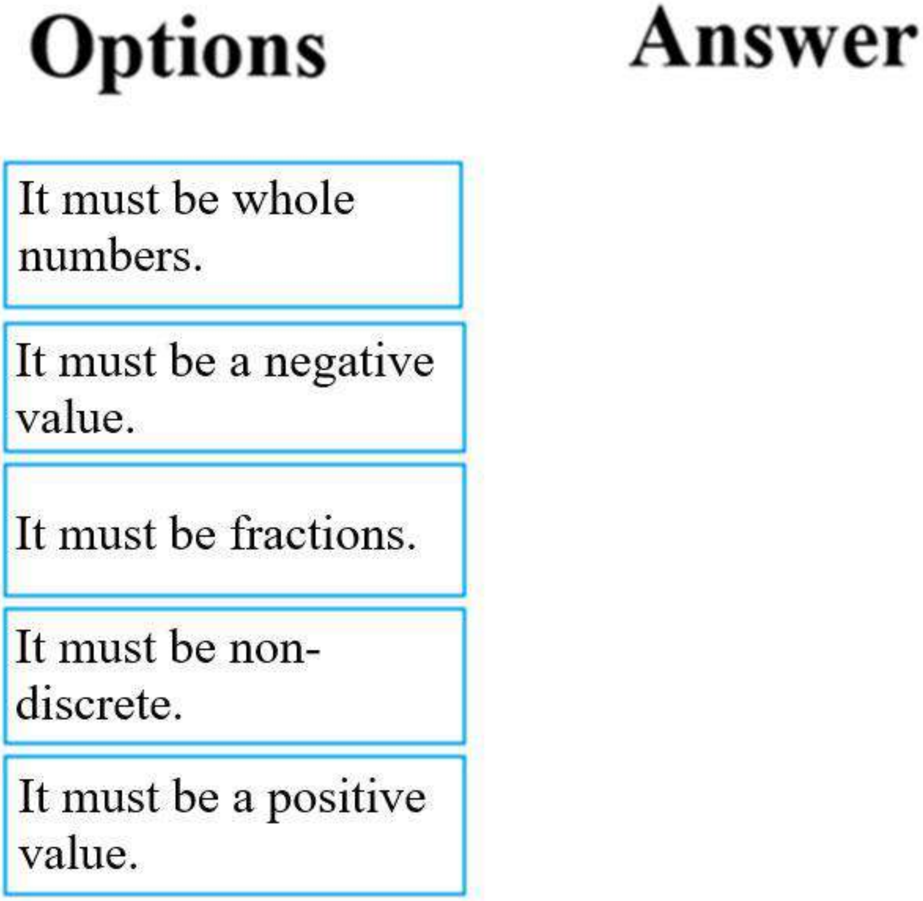
Question #23*Topic 1*
This question is included in a number of questions that depicts the identical set-up. However, every question has a distinctive result. Establish if the recommendation satisfies the requirements.
You have been tasked with evaluating your model on a partial data sample via k-fold cross-validation.
You have already configured a k parameter as the number of splits. You now have to configure the k parameter for the cross-validation with the usual value choice.
Recommendation: You configure the use of the value k=1.
Will the requirements be satisfied?

- A. Yes
- B. No

Correct Answer: *B*

Question #24*Topic 1*
DRAG DROP -
You are in the process of constructing a regression model.
You would like to make it a Poisson regression model. To achieve your goal, the feature values need to meet certain conditions.
Which of the following are relevant conditions with regards to the label data? Answer by dragging the correct options from the list to the answer area.

Select and Place:



Correct Answer:

Options	Answer
It must be whole numbers.	It must be whole numbers.
It must be a negative value.	It must be a positive value.
It must be fractions.	
It must be non-discrete.	
It must be a positive value.	

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.

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

Question #25*Topic 1*

This question is included in a number of questions that depicts the identical set-up. However, every question has a distinctive result. Establish if the recommendation satisfies the requirements.
You are in the process of carrying out feature engineering on a dataset.
You want to add a feature to the dataset and fill the column value.
Recommendation: You must make use of the Group Categorical Values Azure Machine Learning Studio module.
Will the requirements be satisfied?

- A. Yes

- B. No

Correct Answer: *A*

Question #26*Topic 1*
This question is included in a number of questions that depicts the identical set-up. However, every question has a distinctive result. Establish if the recommendation satisfies the requirements.
You are in the process of carrying out feature engineering on a dataset.
You want to add a feature to the dataset and fill the column value.
Recommendation: You must make use of the Join Data Azure Machine Learning Studio module.
Will the requirements be satisfied?

- A. Yes
- B. No

Correct Answer: *B*

Question #27*Topic 1*
This question is included in a number of questions that depicts the identical set-up. However, every question has a distinctive result. Establish if the recommendation satisfies the requirements.
You are in the process of carrying out feature engineering on a dataset.
You want to add a feature to the dataset and fill the column value.
Recommendation: You must make use of the Edit Metadata Azure Machine Learning Studio module.
Will the requirements be satisfied?

- A. Yes
- B. No

Correct Answer: *B*
Typical metadata changes might include marking columns as features.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/edit-metadata
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/join-data
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/group-categorical-values

Question #28*Topic 1*
You have been tasked with ascertaining if two sets of data differ considerably. You will make use of Azure Machine Learning Studio to complete your task.
You plan to perform a paired t-test.
Which of the following are conditions that must apply to use a paired t-test? (Choose all that apply.)

- A. All scores are independent from each other.
- B. You have a matched pairs of scores.
- C. The sampling distribution of d is normal.
- D. The sampling distribution of x1- x2 is normal.

Correct Answer: *BC*
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/test-hypothesis-using-t-test

How to configure Test Hypothesis Using t-Test

Use a single dataset as input. The columns that you are comparing must be in the same dataset.

If you need to compare columns from different datasets, you can isolate each column to compare by using Select Columns in Dataset, and then merge them into one dataset by using Add Columns.

1. Add the Test Hypothesis Using t-Test module to your experiment.

   You can find this module in the Statistical Functions category in Studio (classic).

2. Add the dataset that contains the column or columns that you want to analyze.

3. Decide which kind of t-test is appropriate for your data. See How to choose a t-test.

4. **Single sample**: If you are using a single sample, set these parameters:

Question #29*Topic 1*
You want to train a classification model using data located in a comma-separated values (CSV) file.
The classification model will be trained via the Automated Machine Learning interface using the Classification task type.
You have been informed that only linear models need to be assessed by the Automated Machine Learning.
Which of the following actions should you take?

- A. You should disable deep learning.
- B. You should enable automatic featurization.
- C. You should disable automatic featurization.
- D. You should set the task type to Forecasting.

Correct Answer: *A*
Reference:
https://econml.azurewebsites.net/spec/estimation/dml.html
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-use-automated-ml-for-ml-models

Question #30*Topic 1*
You are preparing to train a regression model via automated machine learning. The data available to you has features with missing values, as well as categorical features with little discrete values.
You want to make sure that automated machine learning is configured as follows:
✑ missing values must be automatically imputed.

⇨ categorical features must be encoded as part of the training task.
Which of the following actions should you take?

- A. You should make use of the featurization parameter with the 'auto' value pair.
- B. You should make use of the featurization parameter with the 'off' value pair.
- C. You should make use of the featurization parameter with the 'on' value pair.
- D. You should make use of the featurization parameter with the 'FeaturizationConfig' value pair.

Correct Answer: *A*
Featurization str or FeaturizationConfig
Values: 'auto' / 'off' / FeaturizationConfig
Indicator for whether featurization step should be done automatically or not, or whether customized featurization should be used.
Column type is automatically detected. Based on the detected column type preprocessing/featurization is done as follows:
Categorical: Target encoding, one hot encoding, drop high cardinality categories, impute missing values.
Numeric: Impute missing values, cluster distance, weight of evidence.
DateTime: Several features such as day, seconds, minutes, hours etc.
Text: Bag of words, pre-trained Word embedding, text target encoding.
Reference:
https://docs.microsoft.com/en-us/python/api/azureml-train-automl-client/azureml.train.automl.automlconfig.automlconfig

Constructor

Python Copy

```python
AutoMLConfig(task: str, path: typing.Union[str, NoneType] = None, iterations:
typing.Union[int, NoneType] = None, primary_metric: typing.Union[str, NoneType] =
None, positive_label: typing.Union[typing.Any, NoneType] = None, compute_target:
typing.Union[typing.Any, NoneType] = None, spark_context:
typing.Union[typing.Any, NoneType] = None, X: typing.Union[typing.Any, NoneType]
= None, y: typing.Union[typing.Any, NoneType] = None, sample_weight:
typing.Union[typing.Any, NoneType] = None, X_valid: typing.Union[typing.Any,
NoneType] = None, y_valid: typing.Union[typing.Any, NoneType] = None,
sample_weight_valid: typing.Union[typing.Any, NoneType] = None,
cv_splits_indices: typing.Union[typing.List[typing.List[typing.Any]], NoneType] =
None, validation_size: typing.Union[float, NoneType] = None, n_cross_validations:
typing.Union[int, NoneType] = None, y_min: typing.Union[float, NoneType] = None,
y_max: typing.Union[float, NoneType] = None, num_classes: typing.Union[int,
NoneType] = None, featurization: typing.Union[str,
azureml.automl.core.featurization.featurizationconfig.FeaturizationConfig] =
'auto', max_cores_per_iteration: int = 1, max_concurrent_iterations: int = 1,
iteration_timeout_minutes: typing.Union[int, NoneType] = None, mem_in_mb:
typing.Union[int, NoneType] = None, enforce_time_on_windows: bool = True
```

Question #31*Topic 1*
You make use of Azure Machine Learning Studio to develop a linear regression model. You perform an experiment to assess various algorithms.

Which of the following is an algorithm that reduces the variances between actual and predicted values?

- A. Fast Forest Quantile Regression
- B. Poisson Regression
- C. Boosted Decision Tree Regression
- D. Linear Regression

Correct Answer: *D*
Mean absolute error (MAE) measures how close the predictions are to the actual outcomes; thus, a lower score is better.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/algorithm-module-reference/boosted-decision-tree-regression https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/evaluate-model https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/linear-regression

Question #32*Topic 1*
This question is included in a number of questions that depicts the identical set-up. However, every question has a distinctive result. Establish if the recommendation satisfies the requirements.
You have been tasked with constructing a machine learning model that translates language text into a different language text.
The machine learning model must be constructed and trained to learn the sequence of the.
Recommendation: You make use of Convolutional Neural Networks (CNNs).
Will the requirements be satisfied?

- A. Yes
- B. No

Correct Answer: *B*

Question #33*Topic 1*
This question is included in a number of questions that depicts the identical set-up. However, every question has a distinctive result. Establish if the recommendation satisfies the requirements.
You have been tasked with constructing a machine learning model that translates language text into a different language text.
The machine learning model must be constructed and trained to learn the sequence of the.
Recommendation: You make use of Generative Adversarial Networks (GANs).
Will the requirements be satisfied?

- A. Yes
- B. No

Correct Answer: *B*

Question #34*Topic 1*
This question is included in a number of questions that depicts the identical set-up. However, every question has a distinctive result. Establish if the recommendation satisfies the requirements.

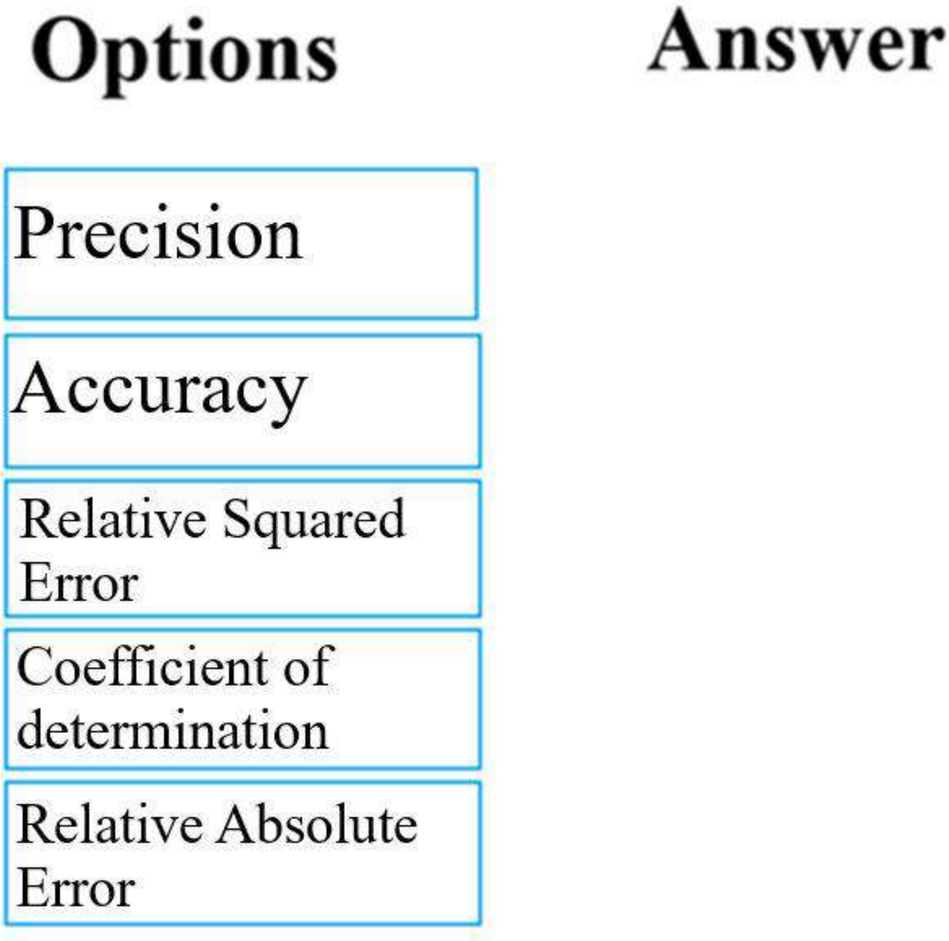
You have been tasked with constructing a machine learning model that translates language text into a different language text.
The machine learning model must be constructed and trained to learn the sequence of the.
Recommendation: You make use of Recurrent Neural Networks (RNNs).
Will the requirements be satisfied?

- A. Yes
- B. No

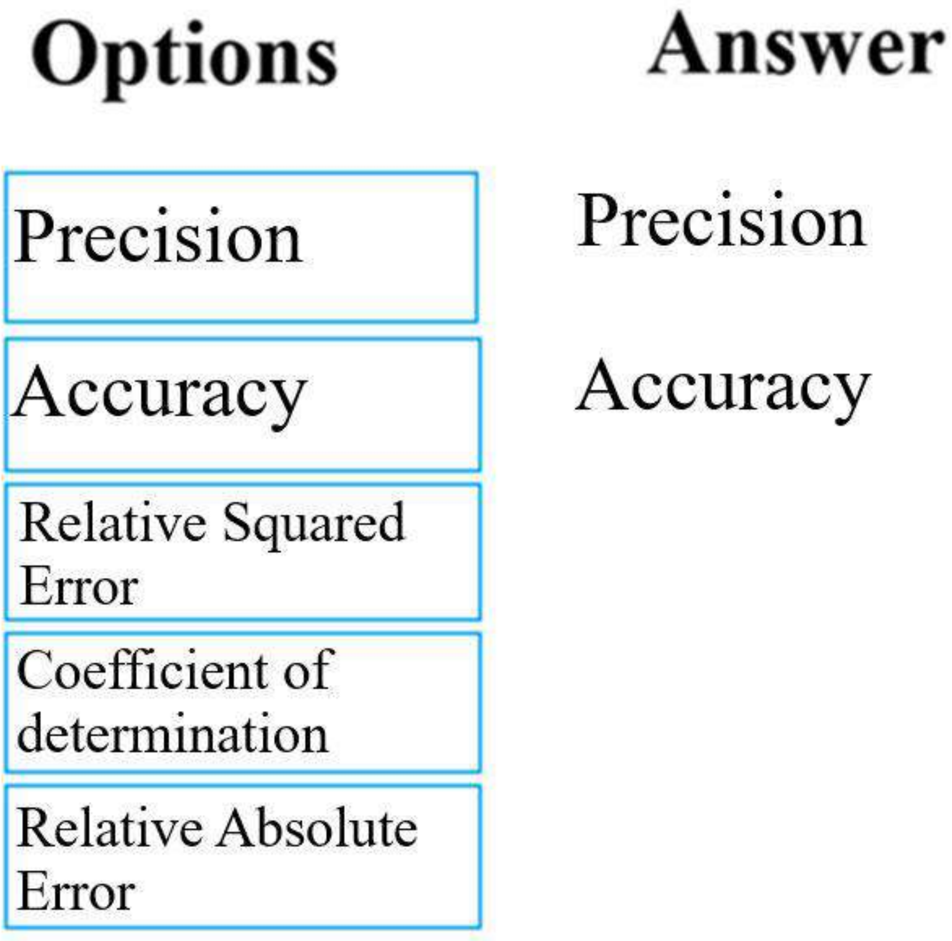
Correct Answer: *A*
Note: RNNs are designed to take sequences of text as inputs or return sequences of text as outputs, or both. They're called recurrent because the network's hidden layers have a loop in which the output and cell state from each time step become inputs at the next time step. This recurrence 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.
Reference:
https://towardsdatascience.com/language-translation-with-rnns-d84d43b40571

Question #35*Topic 1*
DRAG DROP -
You have been tasked with evaluating the performance of a binary classification model that you created.
You need to choose evaluation metrics to achieve your goal.
Which of the following are the metrics you would choose? Answer by dragging the correct options from the list to the answer area.

Select and Place:



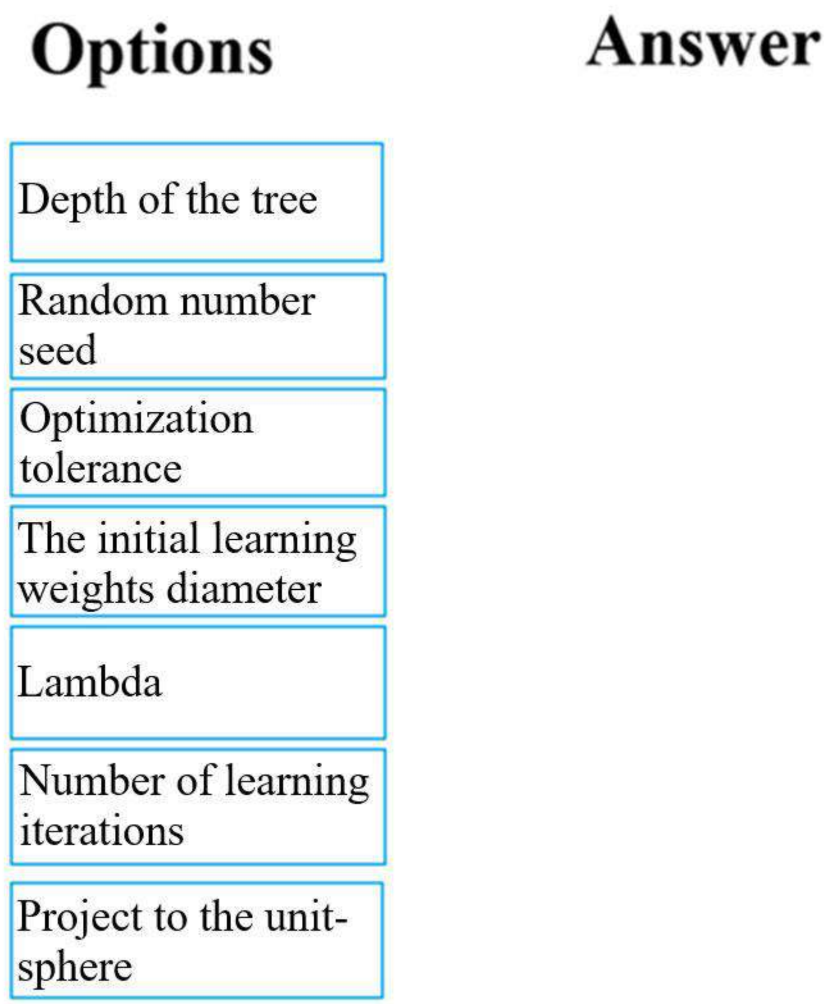
Correct Answer:



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

Question #36*Topic 1*
DRAG DROP -
You build a binary classification model using the Azure Machine Learning Studio Two-Class Neural Network module.
You are preparing to configure the Tune Model Hyperparameters module for the purpose of tuning accuracy for the model.
Which of the following are valid parameters for the Two-Class Neural Network module? Answer by dragging the correct options from the list to the answer area.

Select and Place:



Correct Answer:

Options

- Depth of the tree
- Random number seed
- Optimization tolerance
- The initial learning weights diameter
- Lambda
- Number of learning iterations
- Project to the unit-sphere

Answer

- Random number seed
- The initial learning weights diameter
- Number of learning iterations

Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/two-class-neural-network
Question #37*Topic 1*

You make use of Azure Machine Learning Studio to create a binary classification model.
You are preparing to carry out a parameter sweep of the model to tune hyperparameters. You have to make sure that the sweep allows for every possible combination of hyperparameters to be iterated. Also, the computing resources needed to carry out the sweep must be reduced.
Which of the following actions should you take?

- A. You should consider making use of the Selective grid sweep mode.
- B. You should consider making use of the Measured grid sweep mode.
- C. You should consider making use of the Entire grid sweep mode.
- D. You should consider making use of the Random grid sweep mode.

Correct Answer: *D*
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.
C: Entire grid: When you select this option, the module loops over a grid predefined by the system, to try different combinations and identify the best learner. This option is useful for cases where you don't know what the best parameter settings might be and want to try all possible combination of values.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/tune-model-hyperparameters

Question #38*Topic 1*

You are in the process of constructing a deep convolutional neural network (CNN). The CNN will be used for image classification.
You notice that the CNN model you constructed displays hints of overfitting.
You want to make sure that overfitting is minimized, and that the model is converged to an optimal fit.
Which of the following is TRUE with regards to achieving your goal?

- A. You have to add an additional dense layer with 512 input units, and reduce the amount of training data.
- B. You have to add L1/L2 regularization, and reduce the amount of training data.
- C. You have to reduce the amount of training data and make use of training data augmentation.
- D. You have to add L1/L2 regularization, and make use of training data augmentation.
- E. You have to add an additional dense layer with 512 input units, and add L1/L2 regularization.

Correct Answer: *D*
B: Weight regularization provides an approach to reduce the overfitting of a deep learning neural network model on the training data and improve the performance of the model on new data, such as the holdout test set.
Keras provides a weight regularization API that allows you to add a penalty for weight size to the loss function.
Three different regularizer instances are provided; they are:
✑ L1: Sum of the absolute weights.
✑ L2: Sum of the squared weights.
✑ L1L2: Sum of the absolute and the squared weights.
Because a fully connected layer occupies most of the parameters, it is prone to overfitting. One method to reduce overfitting is dropout. At each training stage, individual nodes are either "dropped out" of the net with probability 1-p or kept with probability p, so that a reduced network is left; incoming and outgoing edges to a dropped-out node are also removed.
By avoiding training all nodes on all training data, dropout decreases overfitting.
Reference:
https://machinelearningmastery.com/how-to-reduce-overfitting-in-deep-learning-with-weight-regularization/ https://en.wikipedia.org/wiki/Convolutional_neural_network

Question #39*Topic 1*

This question is included in a number of questions that depicts the identical set-up. However, every question has a distinctive result. Establish if the recommendation satisfies the requirements.
You are planning to make use of Azure Machine Learning designer to train models.
You need choose a suitable compute type.
Recommendation: You choose Attached compute.
Will the requirements be satisfied?

- A. Yes
- B. No

Correct Answer: *B*
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-create-attach-compute-studio

Question #40*Topic 1*

This question is included in a number of questions that depicts the identical set-up. However, every question has a distinctive result. Establish if the recommendation satisfies the requirements.
You are planning to make use of Azure Machine Learning designer to train models.
You need choose a suitable compute type.
Recommendation: You choose Inference cluster.
Will the requirements be satisfied?

- A. Yes
- B. No

Correct Answer: *B*
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-create-attach-compute-studio

Question #41*Topic 1*

This question is included in a number of questions that depicts the identical set-up. However, every question has a distinctive result. Establish if the recommendation satisfies the requirements.
You are planning to make use of Azure Machine Learning designer to train models.
You need choose a suitable compute type.
Recommendation: You choose Compute cluster.
Will the requirements be satisfied?

- A. Yes
- B. No

Correct Answer: *A*
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-create-attach-compute-studio

Question #42*Topic 1*

You are making use of the Azure Machine Learning to designer construct an experiment.
After dividing a dataset into training and testing sets, you configure the algorithm to be Two-Class Boosted Decision Tree.
You are preparing to ascertain the Area Under the Curve (AUC).
Which of the following is a sequential combination of the models required to achieve your goal?

- A. Train, Score, Evaluate.
- B. Score, Evaluate, Train.
- C. Evaluate, Export Data, Train.
- D. Train, Score, Export Data.

Correct Answer: *A*

Question #43*Topic 1*

DRAG DROP

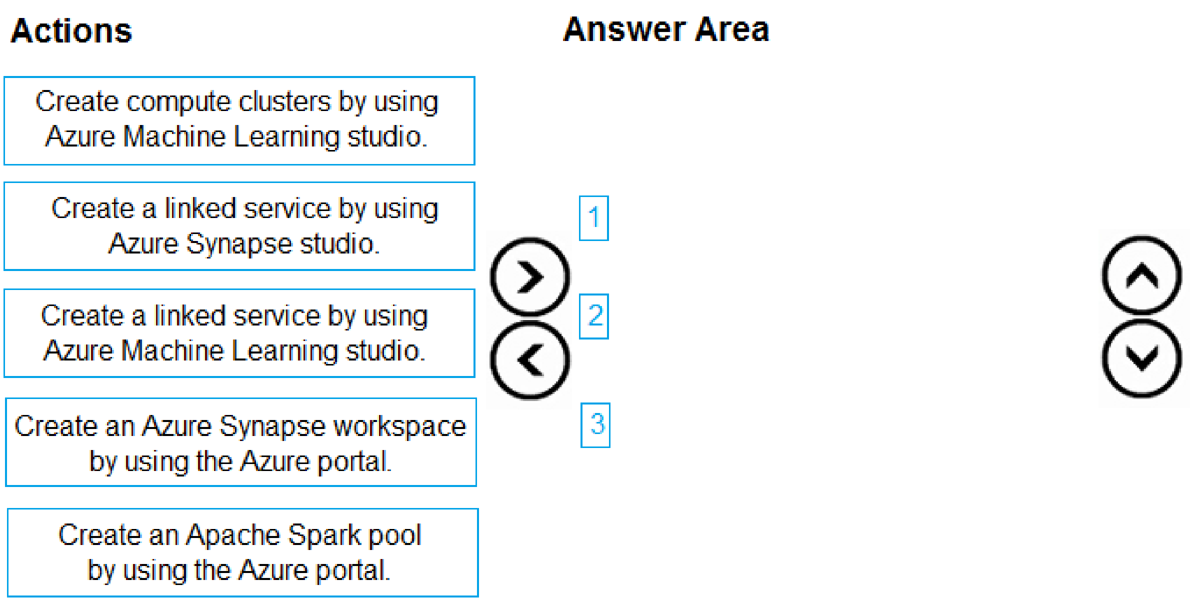
-

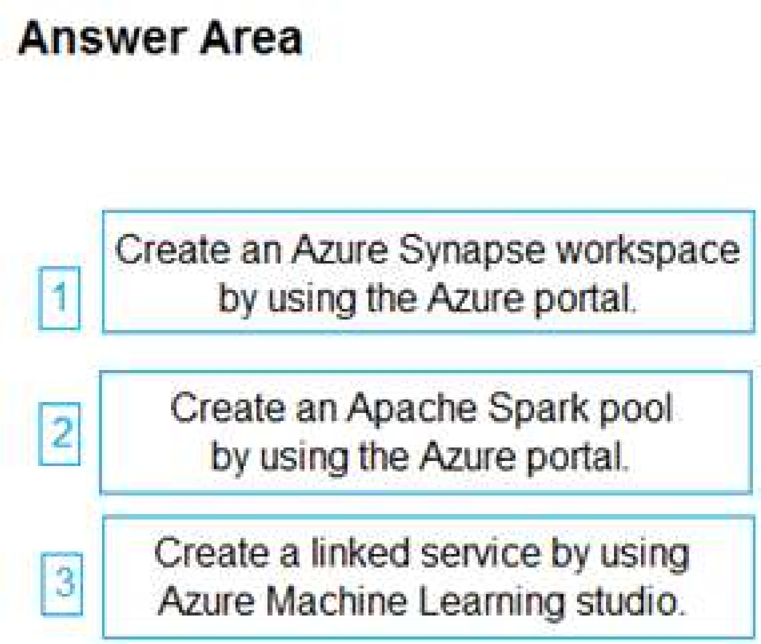
You create an Azure Machine Learning workspace.

You must implement dedicated compute for model training in the workspace by using Azure Synapse compute resources. The solution must attach the dedicated compute and start an Azure Synapse session.

You need to implement the computer resources.

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.





Correct Answer:

Topic 2 - Question Set 2
Question #1*Topic 2*

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

Correct Answer: *CE*
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.
Reference:
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/

Question #2*Topic 2*
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.

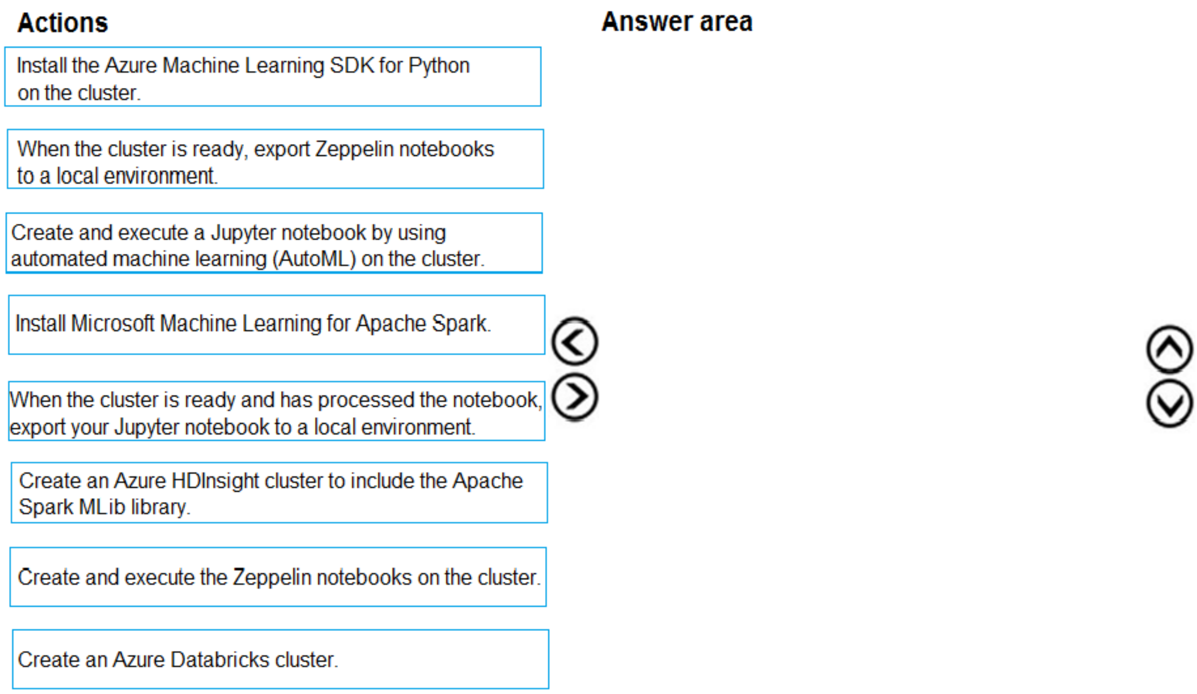
Correct Answer: *B*
In Azure Databricks, we can create two different types of clusters.
⇨ Standard, these are the default clusters and can be used with Python, R, Scala and SQL
⇨ High-concurrency
Azure Databricks is fully integrated with Azure Data Factory.
Incorrect Answers:

D: Azure Container Instances is good for development or testing. Not suitable for production workloads.
Reference:
https://docs.microsoft.com/en-us/azure/architecture/data-guide/technology-choices/data-science-and-machine-learning

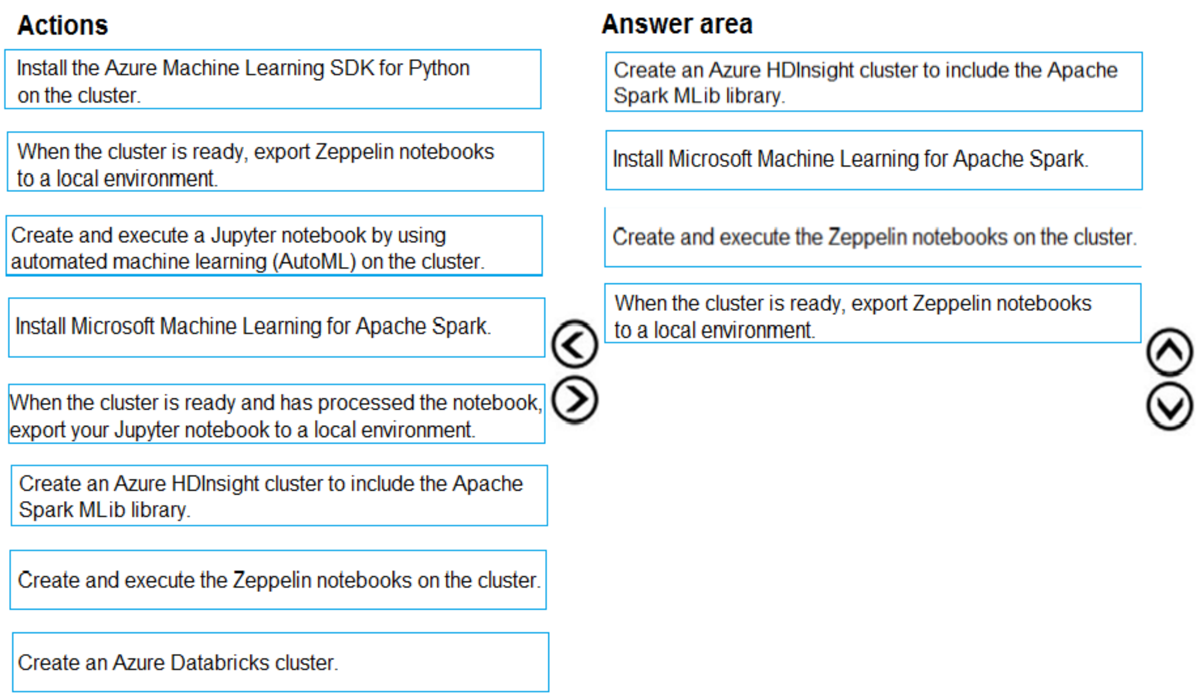
Question #3*Topic 2*

DRAG DROP -
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.
Select and Place:



Correct Answer:



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.
Reference:
https://docs.microsoft.com/en-us/azure/hdinsight/spark/apache-spark-zeppelin-notebook
https://azuremlbuild.blob.core.windows.net/pysparkapi/intro.html
Question #4*Topic 2*

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)

Correct Answer: *A*
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.
Incorrect Answers:
B: Use Machine Learning Studio when you want to experiment with machine learning models quickly and easily, and the built-in machine learning algorithms are sufficient for your solutions.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/data-science-virtual-machine/overview

Question #5*Topic 2*

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 Al 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.

Correct Answer: *C*
In the DSVM, your training models can use deep learning algorithms on hardware that's based on graphics processing units (GPUs).
PostgreSQL is available for the following operating systems: Linux (all recent distributions), 64-bit installers available for macOS (OS X) version 10.6 and newer â€"
Windows (with installers available for 64-bit version; tested on latest versions and back to Windows 2012 R2.
Incorrect Answers:
B: The Azure Geo AI Data Science VM (Geo-DSVM) delivers geospatial analytics capabilities from Microsoft's Data Science VM. Specifically, this VM extends the
AI and data science toolkits in the Data Science VM by adding ESRI's market-leading ArcGIS Pro Geographic Information System.
C, D: DLVM is a template on top of DSVM image. In terms of the packages, GPU drivers etc are all there in the DSVM image. Mostly it is for convenience during creation where we only allow DLVM to be created on GPU VM instances on Azure.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/data-science-virtual-machine/overview

Question #6*Topic 2*

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

Correct Answer: *A*
Azure Cognitive Services expand on Microsoft's evolving portfolio of machine learning APIs and enable developers to easily add cognitive features ג€" such as emotion and video detection; facial, speech, and vision recognition; and speech and language understanding ג€" 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.
Reference:
https://docs.microsoft.com/en-us/azure/cognitive-services/welcome

Question #7*Topic 2*

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)

Correct Answer: *BCD*
You can move data to and from Azure Blob storage using different technologies:
⊸ Azure Storage-Explorer
⊸ AzCopy
⊸ Python
⊸ SSIS
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/move-azure-blob

Question #8*Topic 2*

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

Correct Answer: *C*
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.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/convert-to-arff

Question #9*Topic 2*

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. Rattle
- B. TensorFlow
- C. Weka
- D. Scikit-learn

Correct Answer: *B*
TensorFlow is an open-source library for numerical computation and large-scale machine learning. It uses Python to provide a convenient front-end API for building applications with the framework
TensorFlow can train and run deep neural networks for handwritten digit classification, image recognition, word embeddings, recurrent neural networks, sequence- to-sequence models for machine translation, natural language processing, and PDE (partial differential equation) based simulations.
Incorrect Answers:
A: Rattle is the R analytical tool that gets you started with data analytics and machine learning.
C: Weka is used for visual data mining and machine learning software in Java.
D: Scikit-learn is one of the most useful libraries for machine learning in Python. It is on NumPy, SciPy and matplotlib, this library contains a lot of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction.
Reference:
https://www.infoworld.com/article/3278008/what-is-tensorflow-the-machine-learning-library-explained.html
Question #10*Topic 2*

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 DLVM to support CUDA.
What should you implement?

- A. Solid State Drives (SSD)
- B. Computer Processing Unit (CPU) speed increase by using overclocking

- C. Graphic Processing Unit (GPU)
- D. High Random Access Memory (RAM) configuration
- E. Intel Software Guard Extensions (Intel SGX) technology

Correct Answer: *C*
A Deep Learning Virtual Machine is a pre-configured environment for deep learning using GPU instances.
Reference:
https://azuremarketplace.microsoft.com/en-au/marketplace/apps/microsoft-ads.dsvm-deep-learning

Question #11*Topic 2*

You plan to use a Data Science Virtual Machine (DSVM) with the open source deep learning frameworks Caffe2 and PyTorch.
You need to select a pre-configured DSVM to support the frameworks.
What should you create?

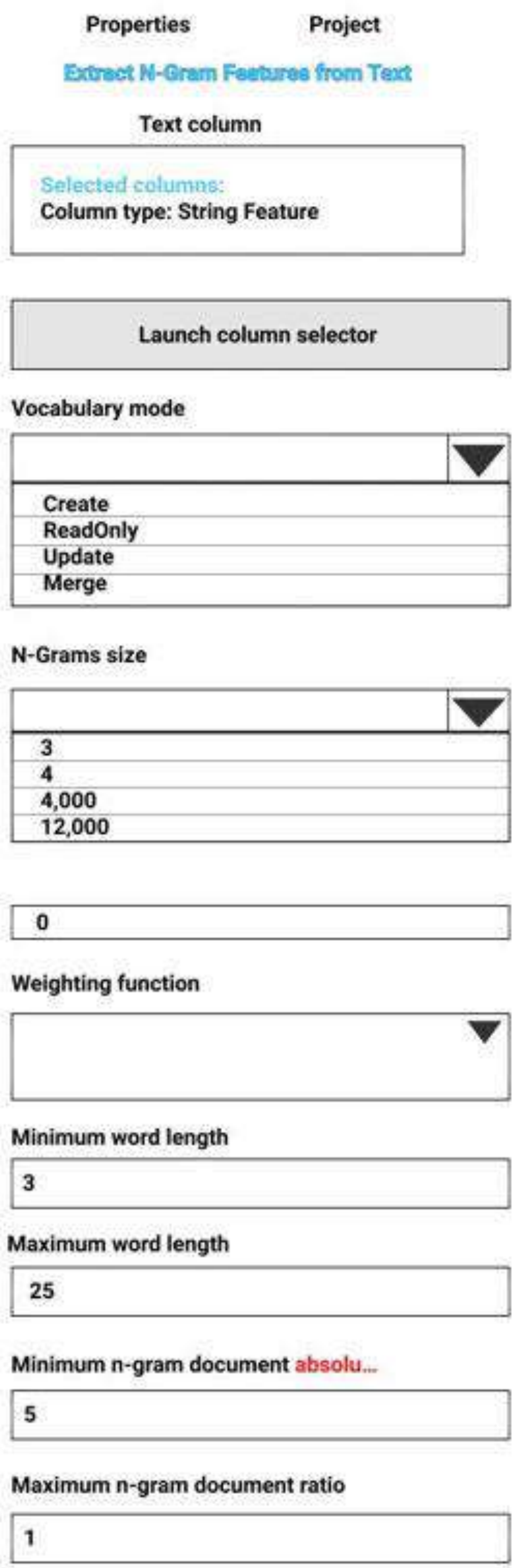
- A. Data Science Virtual Machine for Windows 2012
- B. Data Science Virtual Machine for Linux (CentOS)
- C. Geo AI Data Science Virtual Machine with ArcGIS
- D. Data Science Virtual Machine for Windows 2016
- E. Data Science Virtual Machine for Linux (Ubuntu)

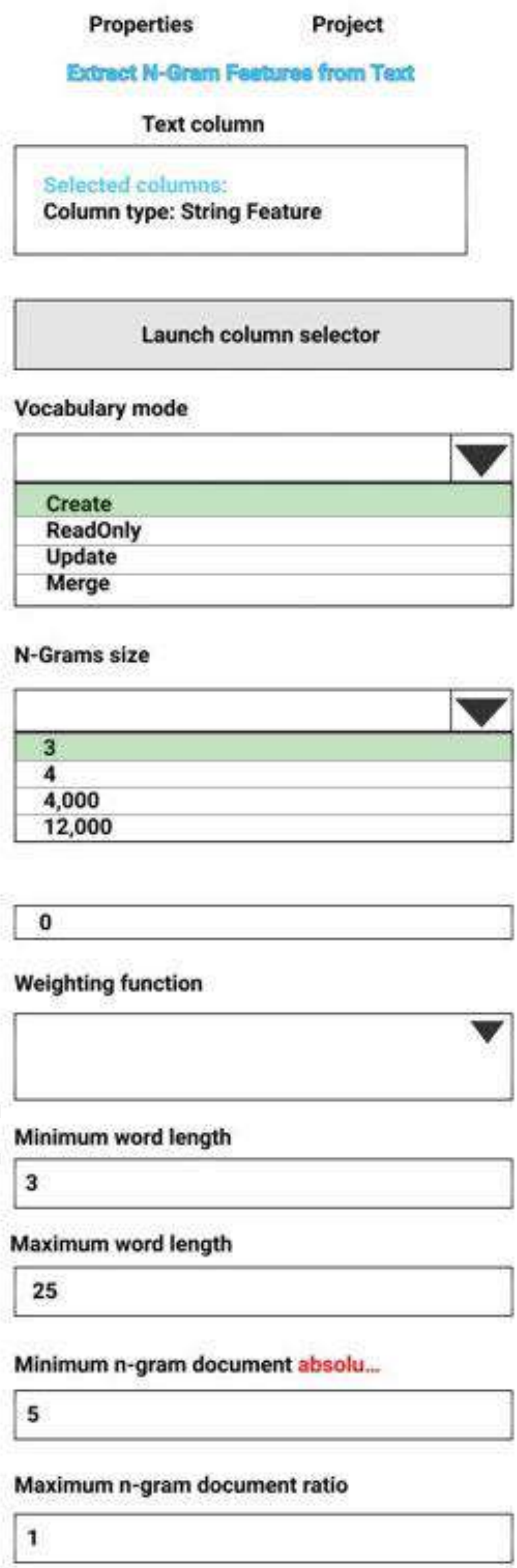
Correct Answer: *E*
Caffe2 and PyTorch is supported by Data Science Virtual Machine for Linux.
Microsoft offers Linux editions of the DSVM on Ubuntu 16.04 LTS and CentOS 7.4.
Only the DSVM on Ubuntu is preconfigured for Caffe2 and PyTorch.
Incorrect Answers:
D: Caffe2 and PytOCH are only supported in the Data Science Virtual Machine for Linux.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/data-science-virtual-machine/overview

Question #12*Topic 2*

HOTSPOT -
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 dictionary from the customer review text and set the maximum n-gram size to trigrams.
What should you select? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.

Hot Area:


Properties
Project
Extract N-Gram Features from Text
Text column
Selected columns:
Column type: String Feature
Launch column selector
Vocabulary mode
Create
ReadOnly
Update
Merge
N-Grams size
3
4
4,000
12,000
0
Weighting function
Minimum word length
3
Maximum word length
25
Minimum n-gram document absolu...
5
Maximum n-gram document ratio
1



Correct Answer:

Vocabulary mode: Create -
For Vocabulary mode, select Create to indicate that you are creating a new list of n-gram features.

N-Grams size: 3 -
For N-Grams size, type a number that indicates the maximum size of the n-grams to extract and store. For example, if you type 3, unigrams, bigrams, and trigrams will be created.

Weighting function: Leave blank -
The option, Weighting function, is required only if you merge or update vocabularies. It specifies how terms in the two vocabularies and their scores should be weighted against each other.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/extract-n-gram-features-from-text
Question #13*Topic 2*

You are developing a data science workspace that uses an Azure Machine Learning service.
You need to select a compute target to deploy the workspace.
What should you use?

- A. Azure Data Lake Analytics
- B. Azure Databricks
- C. Azure Container Service
- D. Apache Spark for HDInsight

Correct Answer: *C*
Azure Container Instances can be used as compute target for testing or development. Use for low-scale CPU-based workloads that require less than 48 GB of
RAM.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/service/how-to-deploy-and-where

Question #14*Topic 2*

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?

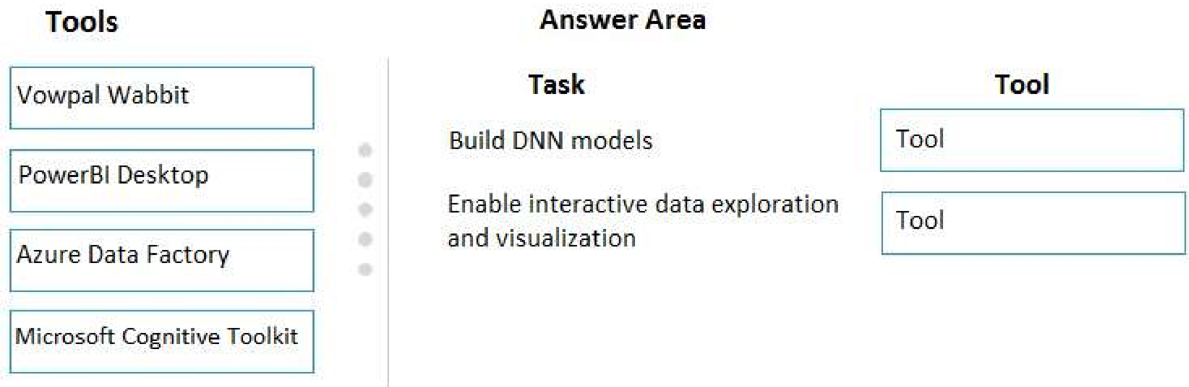
- A. Permutation Feature Importance
- B. Filter Based Feature Selection
- C. Fisher Linear Discriminant Analysis
- D. Synthetic Minority Oversampling Technique (SMOTE)

Correct Answer: *D*
Use the SMOTE module in Azure Machine Learning Studio (classic) to increase the number of underrepresented 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.
You connect the SMOTE module to a dataset that is imbalanced. There are many reasons why a dataset might be imbalanced: the category you are targeting might be very rare in the population, or the data might simply be difficult to collect. Typically, you use SMOTE when the class you want to

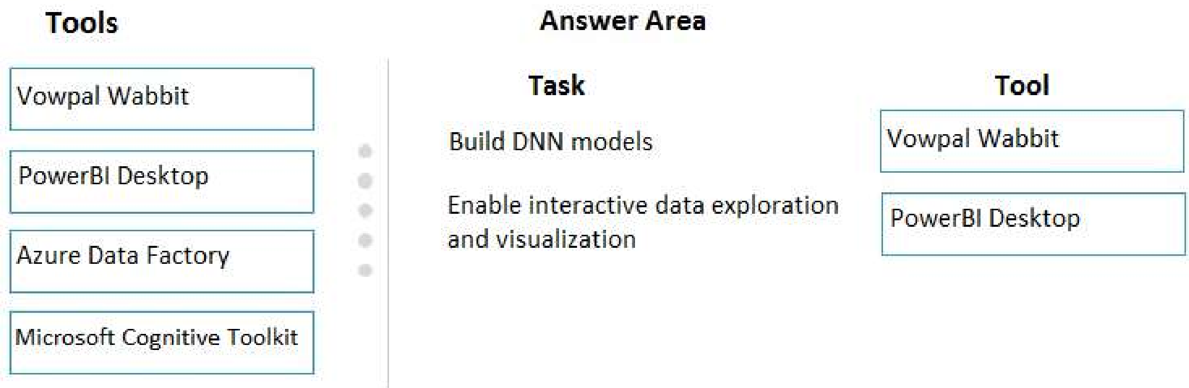
analyze is under- represented.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/smote

Question #15*Topic 2*

DRAG DROP -
You configure a Deep Learning Virtual Machine for Windows.
You need to recommend tools and frameworks to perform the following:
⇨ Build deep neural 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.
Select and Place:



Correct Answer:



Box 1: Vowpal Wabbit -
Use the Train Vowpal Wabbit Version 8 module in Azure Machine Learning Studio (classic), to create a machine learning model by using Vowpal Wabbit.

Box 2: PowerBI Desktop -
Power BI Desktop is a powerful visual data exploration and interactive reporting tool
BI is a name given to a modern approach to business decision making in which users are empowered to find, explore, and share insights from data across the enterprise.

Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/train-vowpal-wabbit-version-8-model https://docs.microsoft.com/en-us/azure/architecture/data-guide/scenarios/interactive-data-exploration
Question #16*Topic 2*

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?

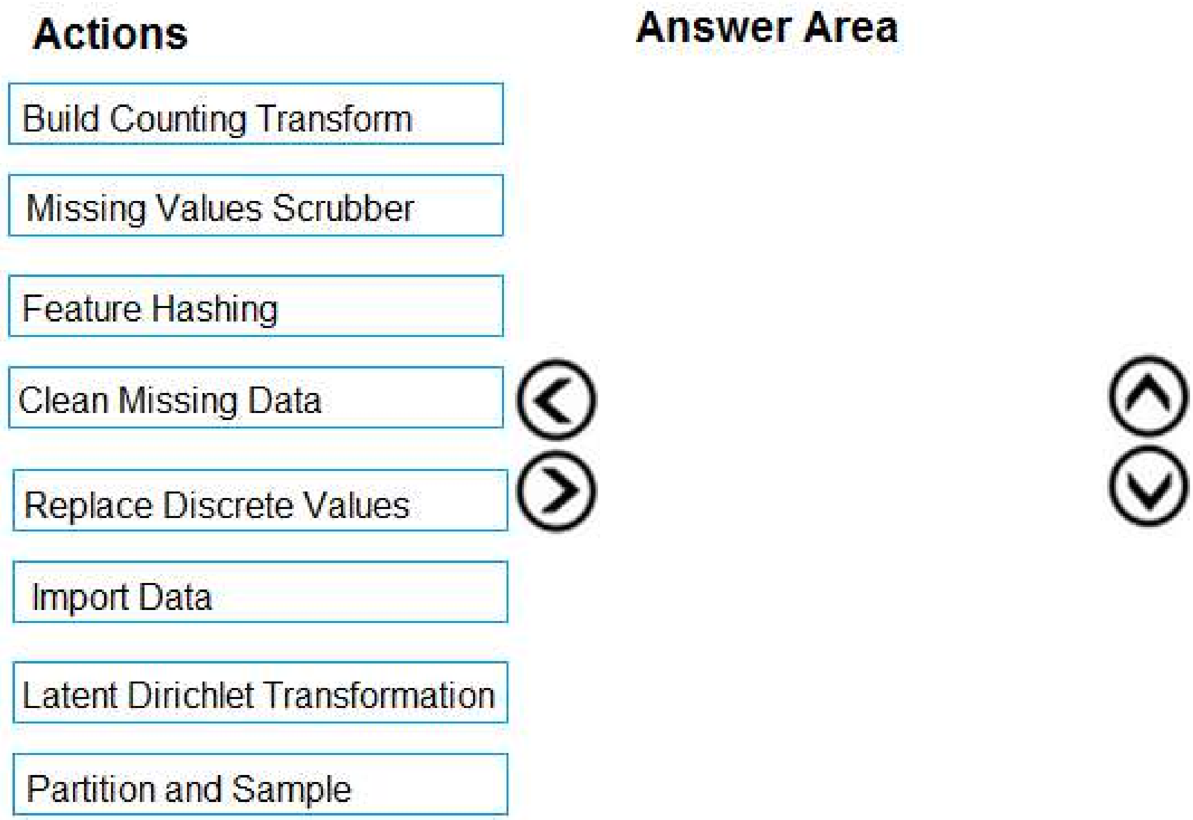
- A. Assign Data to Clusters
- B. Load Trained Model
- C. Partition and Sample
- D. Tune Model-Hyperparameters

Correct Answer: *C*
Partition and Sample with the Stratified split option outputs multiple datasets, partitioned using the rules you specified.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/partition-and-sample

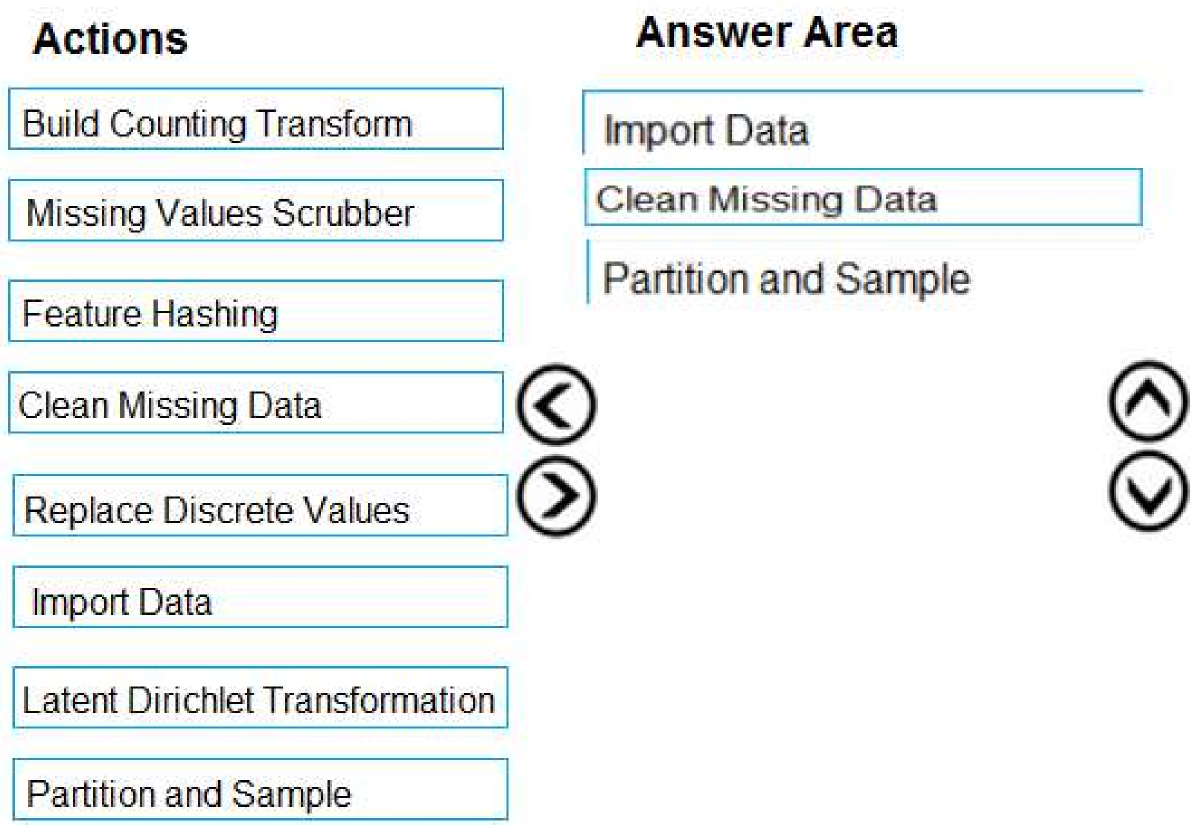
Question #17*Topic 2*

DRAG DROP -
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.

Select and Place:



Correct Answer:



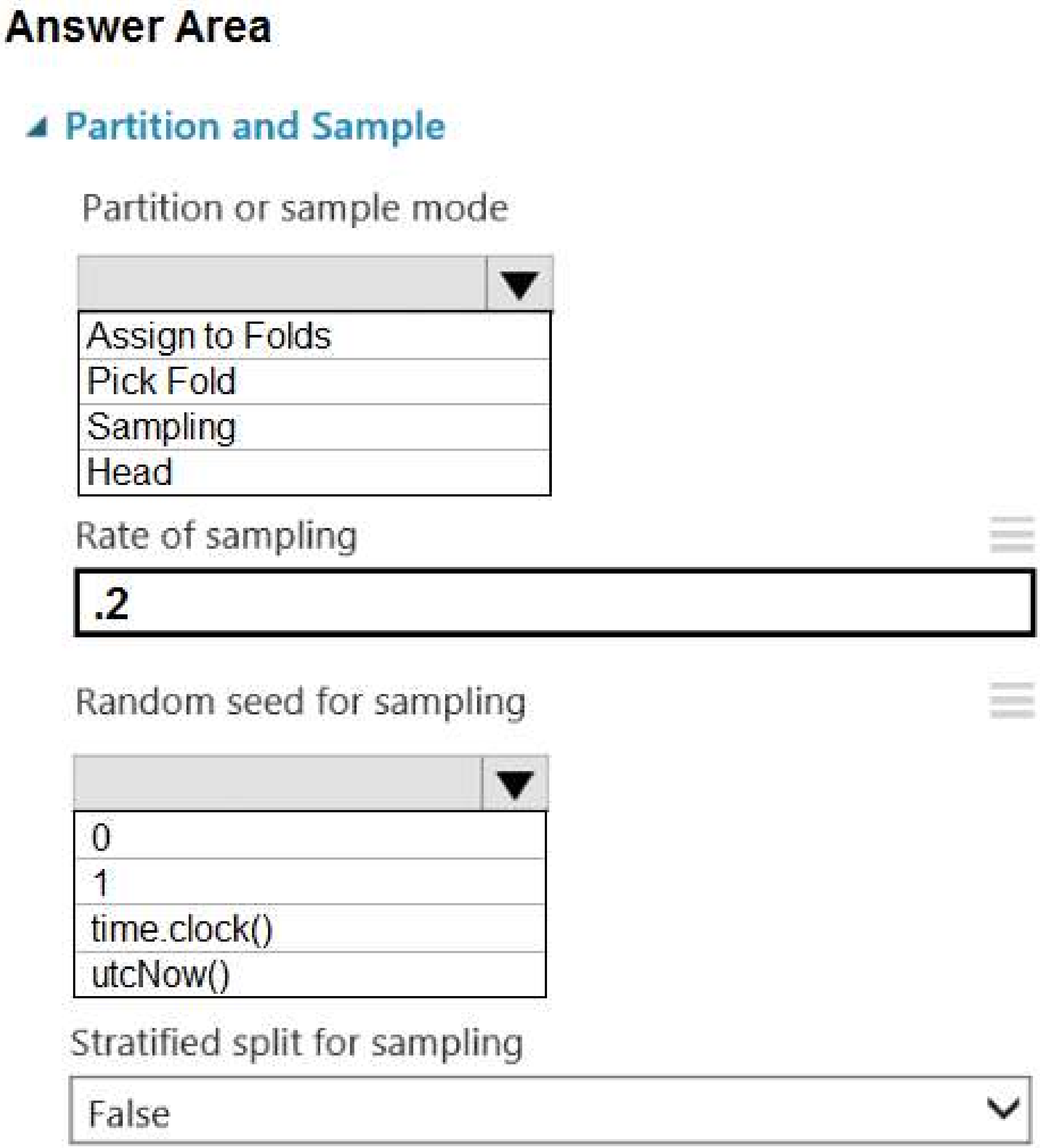
The Clean Missing Data module in Azure Machine Learning Studio, to remove, replace, or infer missing values.
Incorrect Answers:
✑ Latent Direchlet Transformation: Latent Dirichlet Allocation module in Azure Machine Learning Studio, to group otherwise unclassified text into a number of categories. Latent Dirichlet Allocation (LDA) is often used in natural language processing (NLP) to find texts that are similar. Another common term is topic modeling.
✑ Build Counting Transform: Build Counting Transform module in Azure Machine Learning Studio, to analyze training data. From this data, the module builds a count table as well as a set of count-based features that can be used in a predictive model.
Missing Value Scrubber: The Missing Values Scrubber module is deprecated.

▪

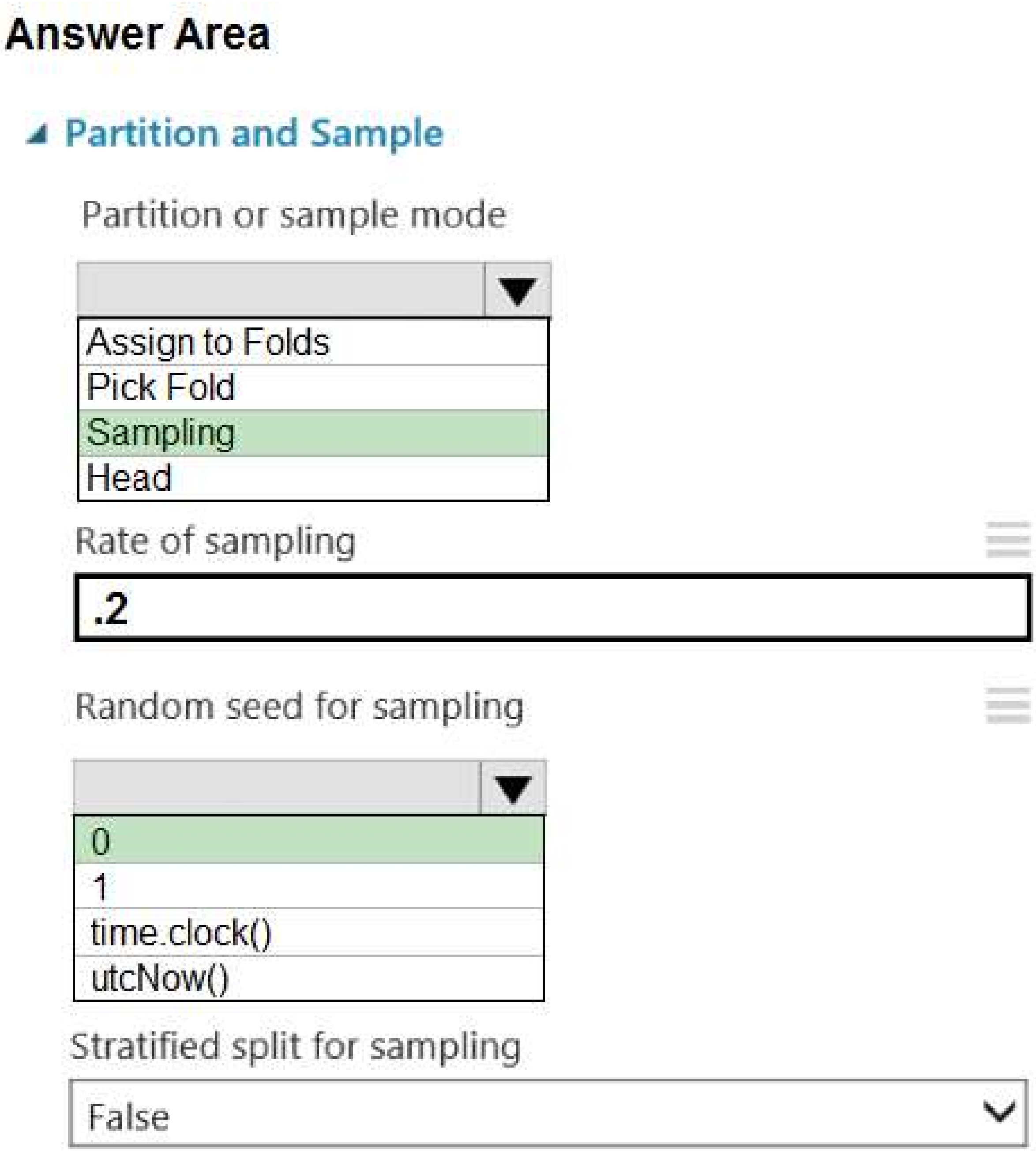
✑ Feature hashing: Feature hashing is used for linguistics, and works by converting unique tokens into integers.
✑ Replace discrete values: the Replace Discrete Values module in Azure Machine Learning Studio is used to generate a probability score that can be used to represent a discrete value. This score can be useful for understanding the information value of the discrete values.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/clean-missing-data

Question #18*Topic 2*
HOTSPOT -
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.

Hot Area:



Correct Answer:



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.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/partition-and-sample
Question #19*Topic 2*

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
- E. Custom substitution value
- F. Replace with mode

Correct Answer: *C*
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.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/clean-missing-data

Question #20*Topic 2*

HOTSPOT -
The finance team asks you to train a model using data in an Azure Storage blob container named finance-data.
You need to register the container as a datastore in an Azure Machine Learning workspace and ensure that an error will be raised if the container does not exist.
How should you complete the code? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.
Hot Area:

Answer Area

```
datastore = Datastore.[ ▼ ](workspace = ws,
```

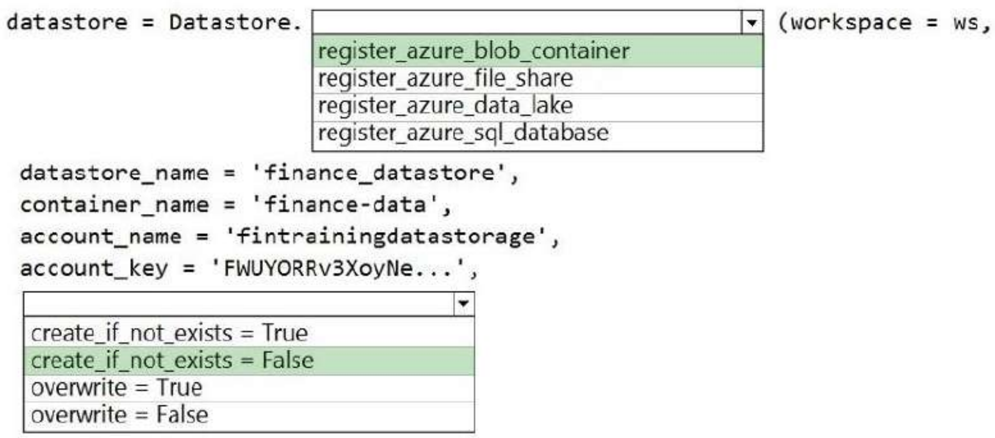
- register_azure_blob_container
- register_azure_file_share
- register_azure_data_lake
- register_azure_sql_database

```
datastore_name = 'finance_datastore',
container_name = 'finance-data',
account_name = 'fintrainingdatastorage',
account_key = 'FWUYORRv3XoyNe...',
[ ▼ ]
```

- create_if_not_exists = True
- create_if_not_exists = False
- overwrite = True
- overwrite = False

Correct Answer:

Answer Area



Box 1: register_azure_blob_container
Register an Azure Blob Container to the datastore.
Box 2: create_if_not_exists = False
Create the file share if it does not exist, defaults to False.
Reference:
https://docs.microsoft.com/en-us/python/api/azureml-core/azureml.core.datastore.datastore
Question #21*Topic 2*

You plan to provision an Azure Machine Learning Basic edition workspace for a data science project.
You need to identify the tasks you will be able to perform in the workspace.
Which three tasks will you be able to perform? Each correct answer presents a complete solution.
NOTE: Each correct selection is worth one point.

- A. Create a Compute Instance and use it to run code in Jupyter notebooks.
- B. Create an Azure Kubernetes Service (AKS) inference cluster.
- C. Use the designer to train a model by dragging and dropping pre-defined modules.
- D. Create a tabular dataset that supports versioning.
- E. Use the Automated Machine Learning user interface to train a model.

Correct Answer: *ABD*
Incorrect Answers:
C, E: The UI is included the Enterprise edition only.
Reference:
https://azure.microsoft.com/en-us/pricing/details/machine-learning/
Question #22*Topic 2*

HOTSPOT -
A coworker registers a datastore in a Machine Learning services workspace by using the following code:

```
Datastore.register_azure_blob_container(workspace=ws,
 datastore_name='demo_datastore',
 container_name='demo_datacontainer',
 account_name='demo_account',
 account_key='0A0A0A-0A0A00A-0A00A0A0A0A0A',
 create_if_not_exists=True)
```

You need to write code to access the datastore from a notebook.
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.
Hot Area:

Answer Area

```
import azureml.core
from azureml.core import Workspace, Datastore
ws = Workspace.from_config()
datastore = [▼] .get( [▼] , ' [▼] ')
```

Box 1	Box 2	Box 3
Workspace	ws	demo_datastore
Datastore	run	demo_datacontainer
Experiment	experiment	demo_account
Run	log	Datastore

Correct Answer:

Answer Area

```
import azureml.core
from azureml.core import Workspace, Datastore
ws = Workspace.from_config()
datastore = [▼] .get( [▼] , ' [▼] ')
```

Box 1	Box 2	Box 3
Workspace	**ws**	**demo_datastore**
Datastore	run	demo_datacontainer
Experiment	experiment	demo_account
Run	log	Datastore

Box 1: DataStore -
To get a specific datastore registered in the current workspace, use the get() static method on the Datastore class:
# Get a named datastore from the current workspace
datastore = Datastore.get(ws, datastore_name='your datastore name')

Box 2: ws -

Box 3: demo_datastore -
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-access-data
Question #23*Topic 2*

A set of CSV files contains sales records. All the CSV files have the same data schema.
Each CSV file contains the sales record for a particular month and has the filename sales.csv. Each

file is stored in a folder that indicates the month and year when the data was recorded. The folders are in an Azure blob container for which a datastore has been defined in an Azure Machine Learning workspace. The folders are organized in a parent folder named sales to create the following hierarchical structure:

```
/sales
  /01-2019
    /sales.csv
  /02-2019
    /sales.csv
  /03-2019
    /sales.csv
 ...
```

At the end of each month, a new folder with that month's sales file is added to the sales folder. You plan to use the sales data to train a machine learning model based on the following requirements:

➾ You must define a dataset that loads all of the sales data to date into a structure that can be easily converted to a dataframe.

➾ You must be able to create experiments that use only data that was created before a specific previous month, ignoring any data that was added after that month.

➾ You must register the minimum number of datasets possible.

You need to register the sales data as a dataset in Azure Machine Learning service workspace. What should you do?

- A. Create a tabular dataset that references the datastore and explicitly specifies each 'sales/mm-yyyy/sales.csv' file every month. Register the dataset with the name sales_dataset each month, replacing the existing dataset and specifying a tag named month indicating the month and year it was registered. Use this dataset for all experiments.
- B. Create a tabular dataset that references the datastore and specifies the path 'sales/*/sales.csv', register the dataset with the name sales_dataset and a tag named month indicating the month and year it was registered, and use this dataset for all experiments.
- C. Create a new tabular dataset that references the datastore and explicitly specifies each 'sales/mm-yyyy/sales.csv' file every month. Register the dataset with the name sales_dataset_MM-YYYY each month with appropriate MM and YYYY values for the month and year. Use the appropriate month-specific dataset for experiments.
- D. Create a tabular dataset that references the datastore and explicitly specifies each 'sales/mm-yyyy/sales.csv' file. Register the dataset with the name sales_dataset each month as a new version and with a tag named month indicating the month and year it was registered. Use this dataset for all experiments, identifying the version to be used based on the month tag as necessary.

Correct Answer: *B*
Specify the path.
Example:
The following code gets the workspace existing workspace and the desired datastore by name. And then passes the datastore and file locations to the path parameter to create a new TabularDataset, weather_ds. from azureml.core import Workspace, Datastore, Dataset datastore_name = 'your datastore name'

```
# get existing workspace
workspace = Workspace.from_config()
# retrieve an existing datastore in the workspace by name
datastore = Datastore.get(workspace, datastore_name)
# create a TabularDataset from 3 file paths in datastore
datastore_paths = [(datastore, 'weather/2018/11.csv'),
(datastore, 'weather/2018/12.csv'),
(datastore, 'weather/2019/*.csv')]
weather_ds = Dataset.Tabular.from_delimited_files(path=datastore_paths)
```

Question #24*Topic 2*

DRAG DROP -

An organization uses Azure Machine Learning service and wants to expand their use of machine learning.

You have the following compute environments. The organization does not want to create another compute environment.

Environment name	Compute type
nb_server	Compute Instance
aks_cluster	Azure Kubernetes Service
mlc_cluster	Machine Learning Compute

You need to determine which compute environment to use for the following scenarios.

Which compute types should you use? To answer, drag the appropriate compute environments to the correct scenarios. Each compute environment 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.

Select and Place:

Environments

- nb_server
- aks_cluster
- mlc_cluster

Answer Area

Scenario	Environment
Run an Azure Machine Learning Designer training pipeline.	Environment
Deploying a web service from the Azure Machine Learning designer.	Environment

Correct Answer:

Environments

- nb_server
- aks_cluster
- mlc_cluster

Answer Area

Scenario	Environment
Run an Azure Machine Learning Designer training pipeline.	nb_server
Deploying a web service from the Azure Machine Learning designer.	mlc_cluster

Box 1: nb_server -

Training targets	Automated ML	ML pipelines	Azure Machine Learning designer
Local computer	yes		
Azure Machine Learning compute cluster	yes & hyperparameter tuning	yes	yes
Azure Machine Learning compute instance	yes & hyperparameter tuning	yes	yes
Remote VM	yes & hyperparameter tuning	yes	
Azure Databricks	yes (SDK local mode only)	yes	
Azure Data Lake Analytics		yes	
Azure HDInsight		yes	
Azure Batch		yes	

Box 2: mlc_cluster -
With Azure Machine Learning, you can train your model on a variety of resources or environments, collectively referred to as compute targets. A compute target can be a local machine or a cloud resource, such as an Azure Machine Learning Compute, Azure HDInsight or a remote virtual machine.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/concept-compute-target
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-set-up-training-targets
Question #25*Topic 2*

HOTSPOT -
You create an Azure Machine Learning compute target named ComputeOne by using the STANDARD_D1 virtual machine image.
ComputeOne is currently idle and has zero active nodes.
You define a Python variable named ws that references the Azure Machine Learning workspace. You run the following Python code:

```
from azureml.core.compute import ComputeTarget, AmlCompute
from azureml.core.compute_target import ComputeTargetException
the_cluster_name = "ComputeOne"
try:
  the_cluster = ComputeTarget(workspace=ws, name=the_cluster_name)
  print('Step1')
except ComputeTargetException:
  config = AmlCompute.provisioning_configuration(vm_size='STANDARD_DS12_v2', max_nodes=4)
  the_cluster = ComputeTarget.create(ws, the_cluster_name, config)
  print('Step2')
```

For each of the following statements, select Yes if the statement is true. Otherwise, select No.
NOTE: Each correct selection is worth one point.
Hot Area:

Answer Area

	Yes	No
A new machine learning compute resource is created with a virtual machine size of STANDARD_DS12_v2 and a maximum of four nodes.	○	○
Any experiments configured to use the_cluster will run on ComputeOne.	○	○
The text **Step1** will be printed to the screen.	○	○

Correct Answer:

Answer Area

	Yes	No
A new machine learning compute resource is created with a virtual machine size of STANDARD_DS12_v2 and a maximum of four nodes.	○	○
Any experiments configured to use the_cluster will run on ComputeOne.	○	○
The text **Step1** will be printed to the screen.	○	○

Box 1: Yes -
ComputeTargetException class: An exception related to failures when creating, interacting with, or configuring a compute target. This exception is commonly raised for failures attaching a compute target, missing headers, and unsupported configuration values.
Create(workspace, name, provisioning_configuration)
Provision a Compute object by specifying a compute type and related configuration.
This method creates a new compute target rather than attaching an existing one.

Box 2: Yes -

Box 3: No -
The line before print('Step1') will fail.
Reference:
https://docs.microsoft.com/en-us/python/api/azureml-core/azureml.core.compute.computetarget
Question #26*Topic 2*

HOTSPOT -
You are developing a deep learning model by using TensorFlow. You plan to run the model training workload on an Azure Machine Learning Compute Instance.
You must use CUDA-based model training.
You need to provision the Compute Instance.
Which two virtual machines sizes can you use? To answer, select the appropriate virtual machine

sizes in the answer area.
NOTE: Each correct selection is worth one point.
Hot Area:

Answer Area

Virtual machine size

Search by name...

Name ↑	vCPUs	GPUs	RAM	Resource disk
BASIC_A0	1		0.75 GB	20 GB
STANDARD_D3_V2	4		14 GB	200 GB
STANDARD_E64_V3	64		432 GB	1,600 GB
STANDARD_M64LS	64		512 GB	2,000 GB
STANDARD_NC12	12	2	112 GB	680 GB
STANDARD_NC24	24	4	224 GB	1,440 GB

Correct Answer:

Answer Area

Virtual machine size

Search by name...

Name ↑	vCPUs	GPUs	RAM	Resource disk
BASIC_A0	1		0.75 GB	20 GB
STANDARD_D3_V2	4		14 GB	200 GB
STANDARD_E64_V3	64		432 GB	1,600 GB
STANDARD_M64LS	64		512 GB	2,000 GB
STANDARD_NC12	12	2	112 GB	680 GB
STANDARD_NC24	24	4	224 GB	1,440 GB

CUDA is a parallel computing platform and programming model developed by Nvidia for general computing on its own GPUs (graphics processing units). CUDA enables developers to speed up compute-intensive applications by harnessing the power of GPUs for the parallelizable part of the computation.
Reference:
https://www.infoworld.com/article/3299703/what-is-cuda-parallel-programming-for-gpus.html
Question #27*Topic 2*

DRAG DROP -
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.

Select and Place:

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		

Correct 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		

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 -
Reference:
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
Question #28*Topic 2*

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 (MDL) binning mode.
Does the solution meet the goal?

- A. Yes
- B. No

Correct Answer: *B*
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.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/group-data-into-bins

Question #29*Topic 2*

HOTSPOT -
You are preparing to use the Azure ML SDK to run an experiment and need to create compute. You run the following code:

```
from azureml.core.compute import ComputeTarget, AmlCompute
from azureml.core.compute_target import ComputeTargetException
ws = Workspace.from_config()
cluster_name = 'aml-cluster'
try:
  training_compute = ComputeTarget(workspace=ws, name=cluster_name)
except ComputeTargetException:
  compute_config = AmlCompute.provisioning_configuration(vm_size='STANDARD_D2_V2', vm_priority='lowpriority',
max_nodes=4)
  training_compute = ComputeTarget.create(ws, cluster_name, compute_config)
  training_compute.wait_for_completion(show_output=True)
```

For each of the following statements, select Yes if the statement is true. Otherwise, select No.
NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

	Yes	No
If a compute cluster named aml-cluster already exists in the workspace, it will be deleted and replaced.	○	○
The `wait_for_completion()` method will not return until the aml-cluster compute has four active nodes.	○	○
If the code creates a new aml-cluster compute target, it may be preempted due to capacity constraints.	○	○
The aml-cluster compute target is deleted from the workspace after the training experiment completes.	○	○

Correct Answer:

Answer Area

	Yes	No
If a compute cluster named aml-cluster already exists in the workspace, it will be deleted and replaced.	○	●
The `wait_for_completion()` method will not return until the aml-cluster compute has four active nodes.	●	○
If the code creates a new aml-cluster compute target, it may be preempted due to capacity constraints.	●	○
The aml-cluster compute target is deleted from the workspace after the training experiment completes.	○	●

Box 1: No -
If a compute cluster already exists it will be used.

Box 2: Yes -
The wait_for_completion method waits for the current provisioning operation to finish on the cluster.

Box 3: Yes -
Low Priority VMs use Azure's excess capacity and are thus cheaper but risk your run being pre-empted.

Box 4: No -
Need to use training_compute.delete() to deprovision and delete the AmlCompute target.
Reference:
https://notebooks.azure.com/azureml/projects/azureml-getting-started/html/how-to-use-azureml/training/train-on-amlcompute/train-on-amlcompute.ipynb https://docs.microsoft.com/en-us/python/api/azureml-core/azureml.core.compute.computetarget
Question #30*Topic 2*

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 normalization with a QuantileIndex normalization.
Does the solution meet the goal?

- A. Yes
- B. No

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

Question #31*Topic 2*

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

Correct Answer: *B*
Instead use the Synthetic Minority Oversampling Technique (SMOTE) sampling mode.
Note: SMOTE is used 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.
Incorrect Answers:
Common data tasks for the Scale and Reduce sampling mode include clipping, binning, and normalizing numerical values.
Reference:
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/data-transformation-scale-and-reduce
Question #32*Topic 2*

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 count for each feature column.
Which two modules can you use? Each correct answer presents a complete solution.
NOTE: Each correct selection is worth one point.

- A. Computer Linear Correlation
- B. Export Count Table
- C. Execute Python Script
- D. Convert to Indicator Values
- E. Summarize Data

Correct Answer: *CE*
The Export Count Table module is provided for backward compatibility with experiments that use the Build Count Table (deprecated) and Count Featurizer
(deprecated) modules.
E: Summarize Data statistics are useful when you want to understand the characteristics of the complete dataset. For example, you might need to know:
➪ How many missing values are there in each column?
➪ How many unique values are there in a feature column?
➪ What is the mean and standard deviation for each column?
➪ The module calculates the important scores for each column, and returns a row of summary statistics for each variable (data column) provided as input.
Incorrect Answers:
A: The Compute Linear Correlation module in Azure Machine Learning Studio is used to compute a set of Pearson correlation coefficients for each possible pair of variables in the input dataset.
C: With Python, you can perform tasks that aren't currently supported by existing Studio modules such as:
Visualizing data using matplotlib
Using Python libraries to enumerate datasets and models in your workspace
Reading, loading, and manipulating data from sources not supported by the Import Data module
D: The purpose of the Convert to Indicator Values 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.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/export-count-table https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/summarize-data

Question #33*Topic 2*

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: Use the Last Observation Carried Forward (LOCF) method to impute the missing data points.
Does the solution meet the goal?

- A. Yes
- B. No

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

Question #34*Topic 2*

HOTSPOT -
You are creating a machine learning model in Python. The provided dataset contains several numerical columns and one text column. The text column represents a product's category. The product category will always be one of the following:
- Bikes
- Cars
- 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.

Hot Area:

Answer Area

```
from sklearn import linear_model
import
```

- pandas as df
- numpy as df
- scipy as df

```
dataset = df.read_csv("data\\ProductSales.csv")
ProductCategoryMapping = {"Bikes":1, "Cars":2, "Boats": 3,
"Vans": 4}
dataset['ProductCategoryMapping'] =
dataset['ProductCategory'].
```

- map[ProductCategoryMapping]
- reduce[ProductCategoryMapping]
- transpose[ProductCategoryMapping]

```
regr = linear_model.LinearRegression()
X_train = dataset[['ProductCategoryMapping', 'ProductSize',
'ProductCost']]
y_train = dataset[['Sales']]
regr.fit(X_train, y_train)
```

Correct Answer:

Answer Area

```
from sklearn import linear_model
import
```

- pandas as df
- numpy as df
- scipy as df

```
dataset = df.read_csv("data\\ProductSales.csv")
ProductCategoryMapping = {"Bikes":1, "Cars":2, "Boats": 3,
"Vans": 4}
dataset['ProductCategoryMapping'] =
dataset['ProductCategory'].
```

- map[ProductCategoryMapping]
- reduce[ProductCategoryMapping]
- transpose[ProductCategoryMapping]

```
regr = linear_model.LinearRegression()
X_train = dataset[['ProductCategoryMapping', 'ProductSize',
'ProductCost']]
y_train = dataset[['Sales']]
regr.fit(X_train, y_train)
```

Box 1: pandas as df -
Pandas takes data (like a CSV or TSV file, or a SQL database) and creates a Python object with rows and columns called data frame that looks very similar to table in a statistical software (think Excel or SPSS for example.
Box 2: transpose[ProductCategoryMapping]
Reshape the data from the pandas Series to columns.
Reference:
https://datascienceplus.com/linear-regression-in-python/
Question #35*Topic 2*

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

Correct Answer: *C*
Reference:
https://notebooks.azure.com/

Question #36*Topic 2*

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

Correct Answer: *A*
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 ג€fully conditional specificationג€ or ג€sequential regression multiple imputationג€ 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.
Reference:
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3074241/
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/clean-missing-data

Question #37*Topic 2*

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

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

Question #38*Topic 2*

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.
You need to select a data cleaning method.
Which method should you use?

- A. Replace using Probabilistic PCA
- B. Normalization
- C. Synthetic Minority Oversampling Technique (SMOTE)
- D. Replace using MICE

Correct Answer: *A*
Replace using Probabilistic PCA: Compared to other options, such as Multiple Imputation using

Chained Equations (MICE), this option has the advantage of not requiring the application of predictors for each column. Instead, it approximates the covariance for the full dataset. Therefore, it might offer better performance for datasets that have missing values in many columns.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/clean-missing-data

Question #39*Topic 2*

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. Split Data
- B. Load Trained Model
- C. Assign Data to Clusters
- D. Group Data into Bins

Correct Answer: *A*
The Group Data into Bins module supports multiple options for binning data. You can customize how the bin edges are set and how values are apportioned into the bins.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/group-data-into-bins

Question #40*Topic 2*

You are a lead data scientist for a project that tracks the health and migration of birds. You create a multi-class image classification deep learning model that uses a set of labeled bird photographs collected by experts.
You have 100,000 photographs of birds. All photographs use the JPG format and are stored in an Azure blob container in an Azure subscription.
You need to access the bird photograph files in the Azure blob container from the Azure Machine Learning service workspace that will be used for deep learning model training. You must minimize data movement.
What should you do?

- A. Create an Azure Data Lake store and move the bird photographs to the store.
- B. Create an Azure Cosmos DB database and attach the Azure Blob containing bird photographs storage to the database.
- C. Create and register a dataset by using TabularDataset class that references the Azure blob storage containing bird photographs.
- D. Register the Azure blob storage containing the bird photographs as a datastore in Azure Machine Learning service.
- E. Copy the bird photographs to the blob datastore that was created with your Azure Machine Learning service workspace.

Correct Answer: *D*
We recommend creating a datastore for an Azure Blob container. When you create a workspace, an Azure blob container and an Azure file share are automatically registered to the workspace.

Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-access-data

Question #41*Topic 2*

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

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

Question #42*Topic 2*

You create an Azure Machine Learning workspace.
You must create a custom role named DataScientist that meets the following requirements:
⇨ Role members must not be able to delete the workspace.
⇨ Role members must not be able to create, update, or delete compute resources in the workspace.
⇨ Role members must not be able to add new users to the workspace.
You need to create a JSON file for the DataScientist role in the Azure Machine Learning workspace.
The custom role must enforce the restrictions specified by the IT Operations team.
Which JSON code segment should you use?
A.

```
{
"Name": "DataScientist",
      "IsCustom": true,
      "Description": "Project Data Scientist role",
      "Actions": ["*"],
      "NotActions":[
      "Microsoft.MachineLearningServices/workspaces/*/delete",
      "Microsoft.MachineLearningServices/workspaces/computes/*/write",
      "Microsoft.MachineLearningServices/workspaces/computes/*/delete",
      "Microsoft.Authorization/*/write"
],
"AssignableScopes": [
  "/subscriptions/<id>/resourceGroups/ml-rg/providers/Microsoft.MachineLearningServices/workspaces/ml-ws"
  ]
}
```

B.

```
{
"Name": "DataScientist",
"IsCustom": true,
"Description": "Project Data Scientist role",
"Actions": ["*"],
"NotActions":[],
"AssignableScopes": [
     "/subscriptions/<id>/resourceGroups/ml-rg/providers/Microsoft.MachineLearningServices/workspaces/ml-ws"
]
}
```

C.

```
{
"Name": "DataScientist",
"IsCustom": true,
"Description": "Project Data Scientist role",
"Actions": ["Microsoft.MachineLearningServices/workspaces/*/delete",
     "Microsoft.MachineLearningServices/workspaces/computes/*/write",
     "Microsoft.MachineLearningServices/workspaces/computes/*/delete",
     "Microsoft.Authorization/*/write"
],
"NotActions":[],
"AssignableScopes":[
     "/subscriptions/<id>/resourceGroups/ml-rg/providers/Microsoft.MachineLearningServices/workspaces/ml-ws"
]
}
```

D.

```
{
"Name": "DataScientist",
"IsCustom": true,
"Description": "Project Data Scientist role",
"Actions": [],
"NotActions":["*"],
"AssignableScopes": [
     "/subscriptions/<id>/resourceGroups/ml-rg/providers/Microsoft.MachineLearningServices/workspaces/ml-ws"
]
}
```

Correct Answer: *A*

The following custom role can do everything in the workspace except for the following actions:

- It can't create or update a compute resource.
- It can't delete a compute resource.
- It can't add, delete, or alter role assignments.
- It can't delete the workspace.

To create a custom role, first construct a role definition JSON file that specifies the permission and scope for the role. The following example defines a custom role named "Data Scientist Custom" scoped at a specific workspace level: data_scientist_custom_role.json :

{
"Name": "Data Scientist Custom",
"IsCustom": true,
"Description": "Can run experiment but can't create or delete compute.",
"Actions": ["*"],
"NotActions": [
"Microsoft.MachineLearningServices/workspaces/*/delete",
"Microsoft.MachineLearningServices/workspaces/write",
"Microsoft.MachineLearningServices/workspaces/computes/*/write",
"Microsoft.MachineLearningServices/workspaces/computes/*/delete",
"Microsoft.Authorization/*/write"
],
"AssignableScopes": [
"/subscriptions/<subscription_id>/resourceGroups/<resource_group_name>/providers/Microsoft.MachineLearningServices/workspaces/

<workspace_name>"
]
}
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-assign-roles
Question #43*Topic 2*

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

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

Question #44*Topic 2*

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

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

Question #45*Topic 2*

HOTSPOT -
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.
Hot Area:

Answer Area

```
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])
```


Correct Answer:

Answer Area

```
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- [ 1 | 2 | 3 | 6 ] ,shuffle - True, random_state-1)

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)
>>> 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]
```

Reference:
https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.KFold.html
Question #46*Topic 2*

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. Recommender Split
- B. Regular Expression Split
- C. Relative Expression Split
- D. Split Rows with the Randomized split parameter set to true

Correct Answer: *C*
Split Rows: Use this option if you just want to divide the data into two parts. You can specify the percentage of data to put in each split, but by default, the data is divided 50-50.
Incorrect Answers:
B: Regular Expression Split: Choose this option when you want to divide your dataset by testing a single column for a value.
C: Relative Expression Split: Use this option whenever you want to apply a condition to a number column.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/split-data

Question #47*Topic 2*

HOTSPOT -
You are preparing to build a deep learning convolutional neural network model for image classification. You create a script to train the model using CUDA devices.
You must submit an experiment that runs this script in the Azure Machine Learning workspace.
The following compute resources are available:
⇨ a Microsoft Surface device on which Microsoft Office has been installed. Corporate IT policies prevent the installation of additional software
⇨ a Compute Instance named ds-workstation in the workspace with 2 CPUs and 8 GB of memory
⇨ an Azure Machine Learning compute target named cpu-cluster with eight CPU-based nodes
⇨ an Azure Machine Learning compute target named gpu-cluster with four CPU and GPU-based nodes
You need to specify the compute resources to be used for running the code to submit the experiment, and for running the script in order to minimize model training time.
Which resources should the data scientist use? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Resource type	Option
Run code to submit the experiment	the Microsoft Surface device the ds-workstation compute instance the cpu-cluster compute target the gpu-cluster compute target
Run the training script	the ds-workstation compute instance the cpu-cluster compute target the gpu-cluster compute target the Microsoft Surface device

Correct Answer:

Answer Area

Resource type	Option
Run code to submit the experiment	the Microsoft Surface device the ds-workstation compute instance the cpu-cluster compute target the gpu-cluster compute target
Run the training script	the ds-workstation compute instance the cpu-cluster compute target the gpu-cluster compute target the Microsoft Surface device

Box 1: the ds-workstation compute instance
A workstation notebook instance is good enough to run experiments.
Box 2: the gpu-cluster compute target
Just as GPUs revolutionized deep learning through unprecedented training and inferencing performance, RAPIDS enables traditional machine learning practitioners to unlock game-changing performance with GPUs. With RAPIDS on Azure Machine Learning service, users can accelerate the entire machine learning pipeline, including data processing, training and inferencing, with GPUs from the NC_v3, NC_v2, ND or ND_v2 families. Users can unlock performance gains of more than 20X (with 4 GPUs), slashing training times from hours to minutes and dramatically reducing time-to-insight.
Reference:
https://azure.microsoft.com/sv-se/blog/azure-machine-learning-service-now-supports-nvidia-s-rapids/
Question #48*Topic 2*

You create an Azure Machine Learning workspace. You are preparing a local Python environment on a laptop computer. You want to use the laptop to connect to the workspace and run experiments.
You create the following config.json file.

```
{
"workspace_name" : "ml-workspace"
}
```

You must use the Azure Machine Learning SDK to interact with data and experiments in the workspace.
You need to configure the config.json file to connect to the workspace from the Python environment.
Which two additional parameters must you add to the config.json file in order to connect to the workspace? Each correct answer presents part of the solution.
NOTE: Each correct selection is worth one point.

- A. login
- B. resource_group
- C. subscription_id
- D. key
- E. region

Correct Answer: *BC*
To use the same workspace in multiple environments, create a JSON configuration file. The configuration file saves your subscription (subscription_id), resource (resource_group), and workspace name so that it can be easily loaded.
The following sample shows how to create a workspace.

```
from azureml.core import Workspace
ws = Workspace.create(name='myworkspace',
subscription_id='<azure-subscription-id>',
resource_group='myresourcegroup',
create_resource_group=True,
location='eastus2'
)
```

Reference:
https://docs.microsoft.com/en-us/python/api/azureml-core/azureml.core.workspace.workspace

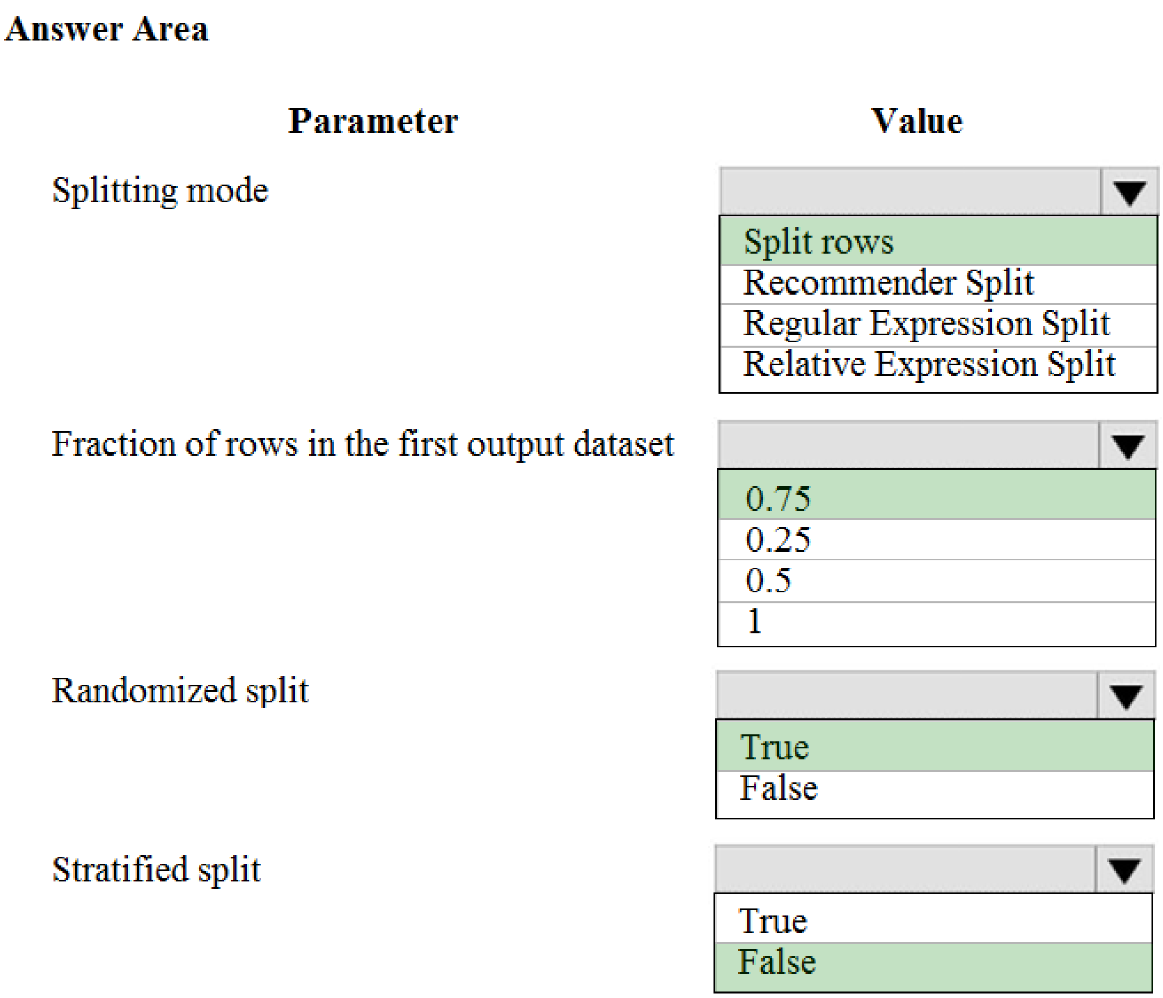
Question #49*Topic 2*

HOTSPOT -
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.
You need to split the data with a 0.75:0.25 ratio.
Which value should you use for each parameter? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.
Hot Area:

Answer Area

Parameter	Value
Splitting mode	Split rows Recommender Split Regular Expression Split Relative Expression Split
Fraction of rows in the first output dataset	0.75 0.25 0.5 1
Randomized split	True False
Stratified split	True False

Correct Answer:



Box 1: Split rows -
Use the Split Rows option if you just want to divide the data into two parts. You can specify the percentage of data to put in each split, but by default, the data is divided 50-50.
You can also randomize the selection of rows in each group, and use stratified sampling. In stratified sampling, you must select a single column of data for which you want values to be apportioned equally among the two result datasets.

Box 2: 0.75 -
If you specify a number as a percentage, or if you use a string that contains the "%" character, the value is interpreted as a percentage. All percentage values must be within the range (0, 100), not including the values 0 and 100.

Box 3: Yes -
To ensure splits are balanced.

Box 4: No -
If you use the option for a stratified split, the output datasets can be further divided by subgroups, by selecting a strata column.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/split-data
Question #50*Topic 2*

You create an Azure Machine Learning compute resource to train models. The compute resource is configured as follows:
✑ Minimum nodes: 2
✑ Maximum nodes: 4
You must decrease the minimum number of nodes and increase the maximum number of nodes to the following values:
✑ Minimum nodes: 0
✑ Maximum nodes: 8
You need to reconfigure the compute resource.
What are three possible ways to achieve this goal? Each correct answer presents a complete solution.
NOTE: Each correct selection is worth one point.

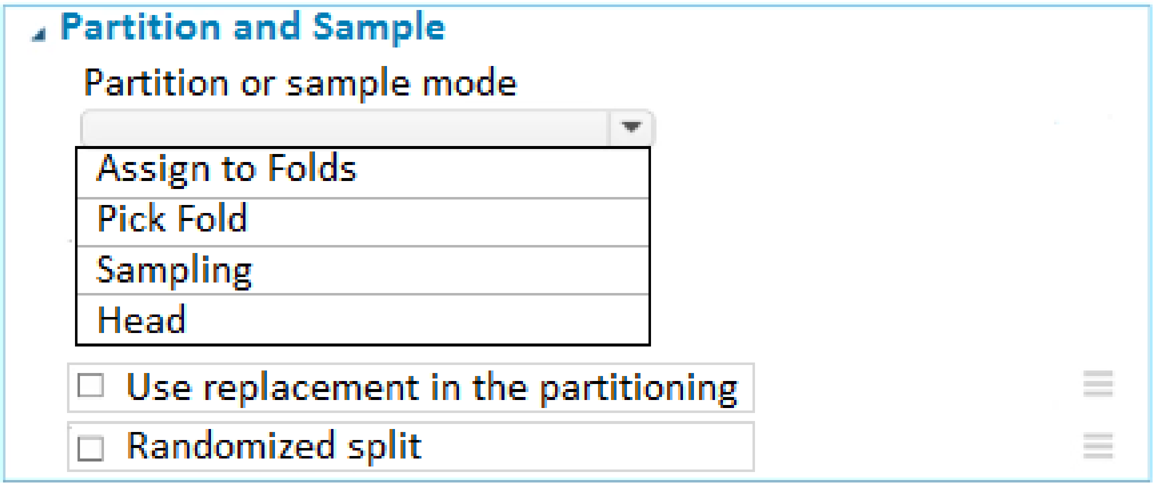
- A. Use the Azure Machine Learning studio.
- B. Run the update method of the AmlCompute class in the Python SDK.
- C. Use the Azure portal.
- D. Use the Azure Machine Learning designer.
- E. Run the refresh_state() method of the BatchCompute class in the Python SDK.

Correct Answer: *ABC*
A: You can manage assets and resources in the Azure Machine Learning studio.
B: The update(min_nodes=None, max_nodes=None, idle_seconds_before_scaledown=None) of the AmlCompute class updates the ScaleSettings for this
AmlCompute target.
C: To change the nodes in the cluster, use the UI for your cluster in the Azure portal.
Reference:
https://docs.microsoft.com/en-us/python/api/azureml-core/azureml.core.compute.amlcompute(class)

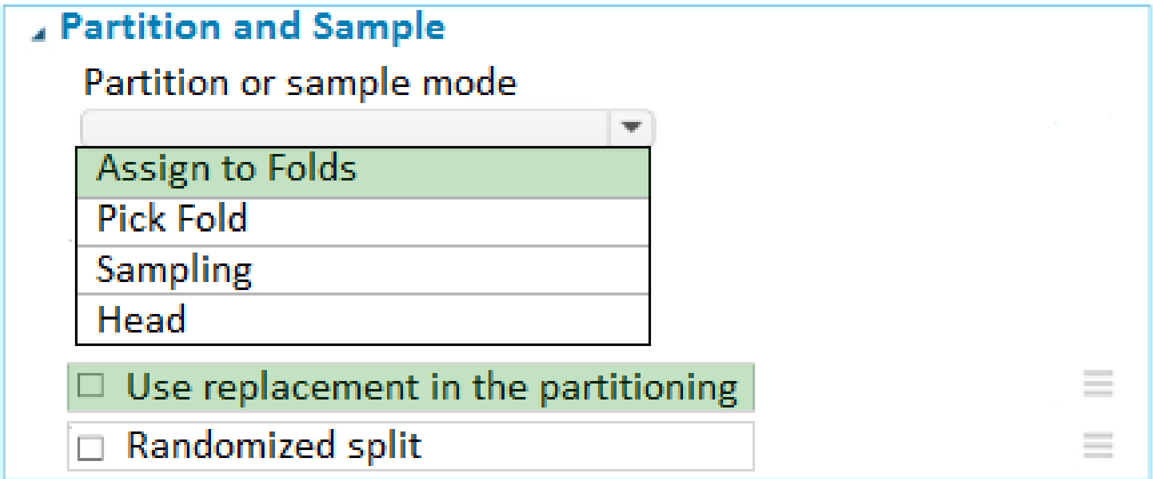
Question #51*Topic 2*

HOTSPOT -
You have a dataset that contains 2,000 rows. You are building a machine learning classification model by using Azure 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 in the dialog box in the answer area.
NOTE: Each correct selection is worth one point.

Hot Area:



Correct Answer:



Use the Split data into partitions option when you want to divide the dataset into subsets of the data. This option is also useful when you want to create a custom number of folds for cross-validation, or to split rows into several groups.

1. Add the Partition and Sample module to your experiment in Studio (classic), and connect the dataset.
2. For Partition or sample mode, select Assign to Folds.
3. Use replacement in the partitioning: Select this option if you want the sampled row to be put back into the pool of rows for potential reuse. As a result, the same row might be assigned to several folds.
4. If you do not use replacement (the default option), the sampled row is not put back into the pool of rows for potential reuse. As a result, each row can be assigned to only one fold.
5. Randomized split: Select this option if you want rows to be randomly assigned to folds.

If you do not select this option, rows are assigned to folds using the round-robin method.

Reference:

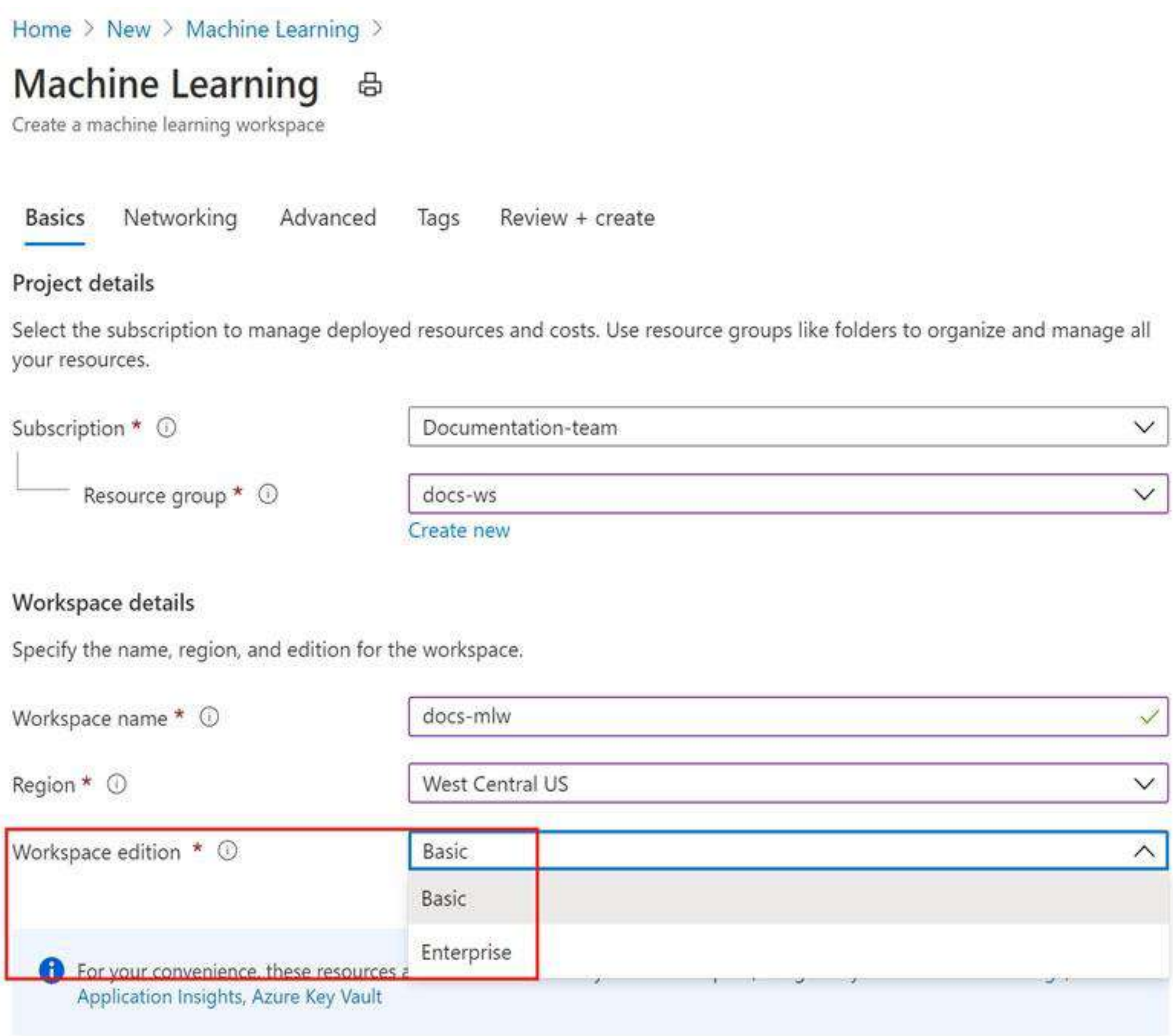
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/partition-and-sample

Question #52*Topic 2*

You create a new Azure subscription. No resources are provisioned in the subscription.
You need to create an Azure Machine Learning workspace.
What are three possible ways to achieve this goal? Each correct answer presents a complete solution.
NOTE: Each correct selection is worth one point.

- A. Run Python code that uses the Azure ML SDK library and calls the Workspace.create method with name, subscription_id, and resource_group parameters.
- B. Navigate to Azure Machine Learning studio and create a workspace.
- C. Use the Azure Command Line Interface (CLI) with the Azure Machine Learning extension to call the az group create function with --name and --location parameters, and then the az ml workspace create function, specifying ג€"w and ג€"g parameters for the workspace name and resource group.
- D. Navigate to Azure Machine Learning studio and create a workspace.
- E. Run Python code that uses the Azure ML SDK library and calls the Workspace.get method with name, subscription_id, and resource_group parameters.

Correct Answer: *ABC*
B: You can create a workspace in the Azure Machine Learning studio
C: You can create a workspace for Azure Machine Learning with Azure CLI
Install the machine learning extension.
Create a resource group: az group create --name <resource-group-name> --location <location>
To create a new workspace where the services are automatically created, use the following command: az ml workspace create -w <workspace-name> -g
<resource-group-name>
D: You can create and manage Azure Machine Learning workspaces in the Azure portal.
1. Sign in to the Azure portal by using the credentials for your Azure subscription.
2. In the upper-left corner of Azure portal, select + Create a resource.
3. Use the search bar to find Machine Learning.
4. Select Machine Learning.
5. In the Machine Learning pane, select Create to begin.



Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-create-workspace-template
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-manage-workspace-cli
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-manage-workspace

Question #53*Topic 2*

HOTSPOT -
You create an Azure Machine Learning workspace and set up a development environment. You plan to train a deep neural network (DNN) by using the
Tensorflow framework and by using estimators to submit training scripts.
You must optimize computation speed for training runs.
You need to choose the appropriate estimator to use as well as the appropriate training compute target configuration.
Which values should you use? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Parameter	Value
Estimator	Estimator SKLearn PyTorch Tensorflow Chainer
Training compute	12 vCPU, 48 GB memory, 96 GB SSD 12 vCPU, 112 GB memory, 680 GB SSD, 2 GPU, 24 GB GPU memory 16 vCPU, 128 GB memory, 160 GB HDD, 80 GB NVME disk (4000 MBps) 44 vCPU, 352 GB memory, 3.4 GHz CPU frequency all cores

Correct Answer:

Answer Area

Parameter	Value
Estimator	Estimator SKLearn PyTorch Tensorflow Chainer
Training compute	12 vCPU, 48 GB memory, 96 GB SSD 12 vCPU, 112 GB memory, 680 GB SSD, 2 GPU, 24 GB GPU memory 16 vCPU, 128 GB memory, 160 GB HDD, 80 GB NVME disk (4000 MBps) 44 vCPU, 352 GB memory, 3.4 GHz CPU frequency all cores

Box 1: Tensorflow -
TensorFlow represents an estimator for training in TensorFlow experiments.
Box 2: 12 vCPU, 112 GB memory..,2 GPU,..
Use GPUs for the deep neural network.
Reference:
https://docs.microsoft.com/en-us/python/api/azureml-train-core/azureml.train.dnn
Question #54*Topic 2*

HOTSPOT -
You have an Azure Machine Learning workspace named workspace1 that is accessible from a public endpoint. The workspace contains an Azure Blob storage datastore named store1 that represents a blob container in an Azure storage account named account1. You configure workspace1 and

account1 to be accessible by using private endpoints in the same virtual network.
You must be able to access the contents of store1 by using the Azure Machine Learning SDK for Python. You must be able to preview the contents of store1 by using Azure Machine Learning studio.
You need to configure store1.
What should you do? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.
Hot Area:

Answer Area

Requirement	Action
Access the contents of store1 by using the Azure Machine Learning SDK for Python.	Set store1 as the default datastore. Disable data validation for store1. Update authentication for store1. Regenerate the keys of account1.
Preview the contents of store1 by using Azure Machine Learning studio.	Set store1 as the default datastore. Disable data validation for store1. Update authentication for store1. Regenerate the keys of account1.

Correct Answer:

Answer Area

Requirement	Action
Access the contents of store1 by using the Azure Machine Learning SDK for Python.	Set store1 as the default datastore. Disable data validation for store1. Update authentication for store1. Regenerate the keys of account1.
Preview the contents of store1 by using Azure Machine Learning studio.	Set store1 as the default datastore. Disable data validation for store1. Update authentication for store1. Regenerate the keys of account1.

Box 1: Regenerate the keys of account1.
Azure Blob Storage support authentication through Account key or SAS token.
To authenticate your access to the underlying storage service, you can provide either your account key, shared access signatures (SAS) tokens, or service principal
Box 2: Update the authentication for store1.
For Azure Machine Learning studio users, several features rely on the ability to read data from a dataset; such as dataset previews, profiles and automated machine learning. For these features to work with storage behind virtual networks, use a workspace managed identity in the studio to allow Azure Machine
Learning to access the storage account from outside the virtual network.
Note: Some of the studio's features are disabled by default in a virtual network. To re-enable these features, you must enable managed identity for storage accounts you intend to use in the studio.
The following operations are disabled by default in a virtual network:

- Preview data in the studio.

Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-access-data

Question #55*Topic 2*

HOTSPOT -

You are using an Azure Machine Learning workspace. You set up an environment for model testing and an environment for production.

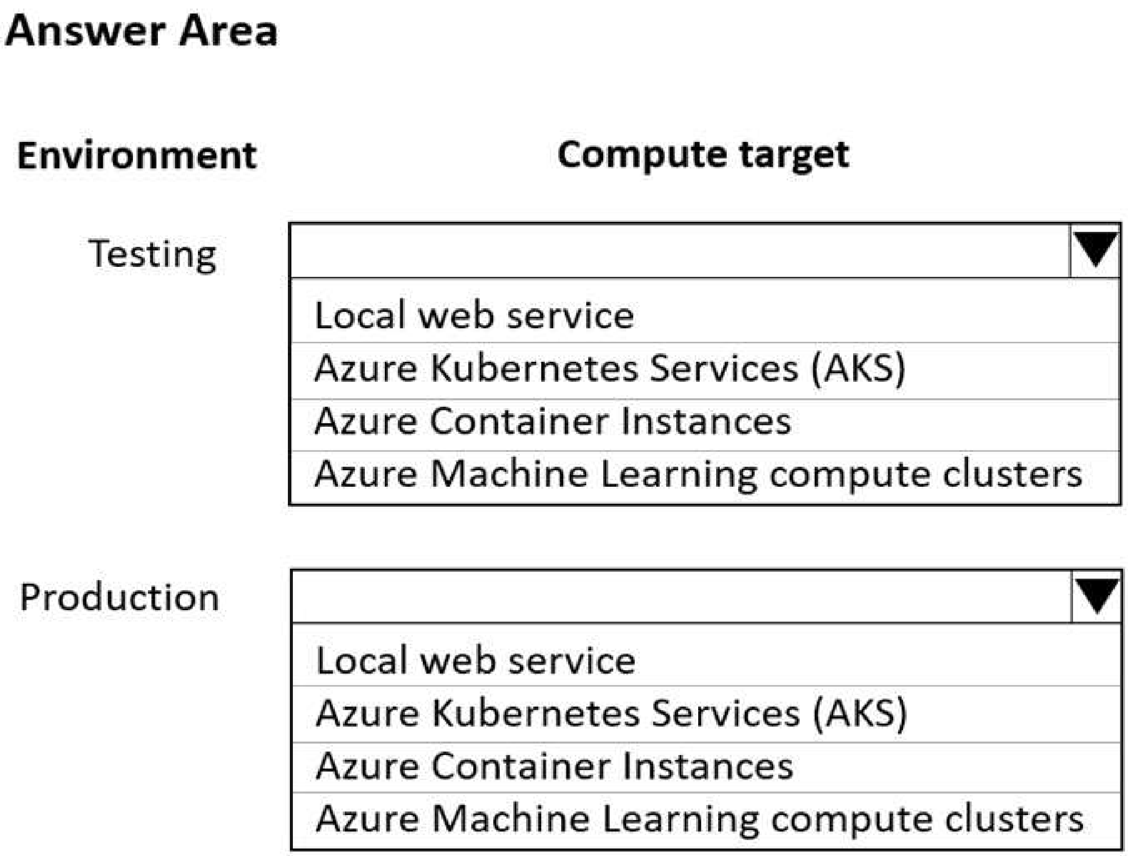
The compute target for testing must minimize cost and deployment efforts. The compute target for production must provide fast response time, autoscaling of the deployed service, and support real-time inferencing.

You need to configure compute targets for model testing and production.

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

NOTE: Each correct selection is worth one point.

Hot Area:



Correct Answer:

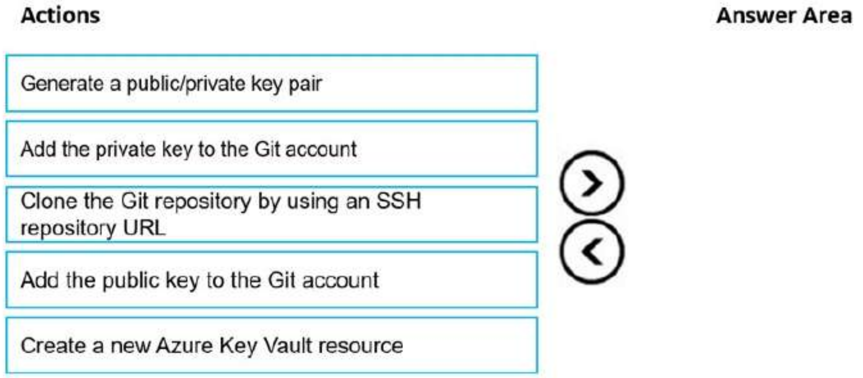
Answer Area

Environment	Compute target
Testing	Local web service (selected) Azure Kubernetes Services (AKS) Azure Container Instances Azure Machine Learning compute clusters
Production	Local web service Azure Kubernetes Services (AKS) (selected) Azure Container Instances Azure Machine Learning compute clusters

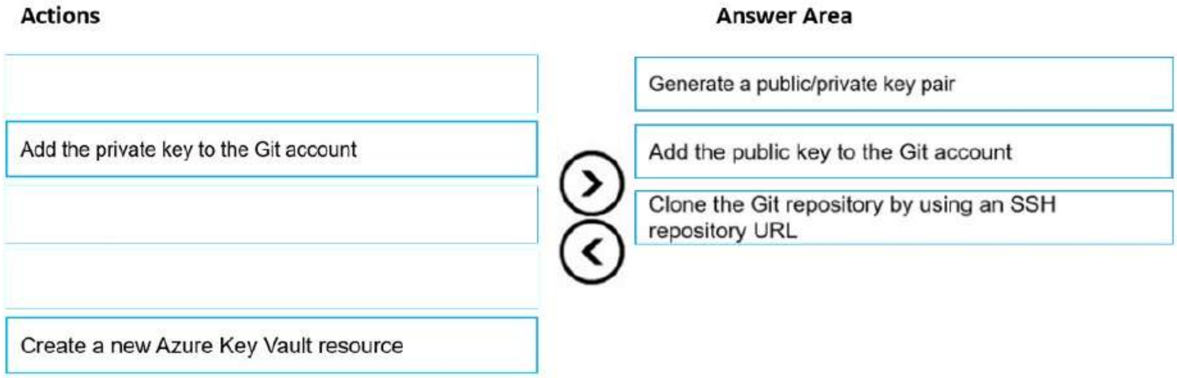
Box 1: Local web service -
The Local web service compute target is used for testing/debugging. Use it for limited testing and troubleshooting. Hardware acceleration depends on use of libraries in the local system.
Box 2: Azure Kubernetes Service (AKS)
Azure Kubernetes Service (AKS) is used for Real-time inference.
Recommended for production workloads.
Use it for high-scale production deployments. Provides fast response time and autoscaling of the deployed service
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/concept-compute-target
Question #56*Topic 2*

DRAG DROP -
You are using a Git repository to track work in an Azure Machine Learning workspace.
You need to authenticate a Git account by using SSH.
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.

Select and Place:



Correct Answer:



Authenticate your Git Account with SSH:
Step 1: Generating a public/private key pair

Generate a new SSH key -
1. Open the terminal window in the Azure Machine Learning Notebook Tab.
2. Paste the text below, substituting in your email address.
ssh-keygen -t rsa -b 4096 -C "your_email@example.com"
This creates a new ssh key, using the provided email as a label.
> Generating public/private rsa key pair.
Step 2: Add the public key to the Git Account
In your terminal window, copy the contents of your public key file.
Step 3: Clone the Git repository by using an SSH repository URL
1. Copy the SSH Git clone URL from the Git repo.
2. Paste the url into the git clone command below, to use your SSH Git repo URL. This will look something like: git clone git@example.com:GitUser/azureml-example.git
Cloning into 'azureml-example'.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/concept-train-model-git-integration
Question #57*Topic 2*

You use Azure Machine Learning to train a model based on a dataset named dataset1.
You define a dataset monitor and create a dataset named dataset2 that contains new data.
You need to compare dataset1 and dataset2 by using the Azure Machine Learning SDK for Python.
Which method of the DataDriftDetector class should you use?

- A. run
- B. get
- C. backfill
- D. update

Correct Answer: *C*
A backfill run is used to see how data changes over time.
Reference:
https://docs.microsoft.com/en-us/python/api/azureml-datadrift/azureml.datadrift.datadriftdetector.datadriftdetector

Question #58*Topic 2*

You use an Azure Machine Learning workspace.
You have a trained model that must be deployed as a web service. Users must authenticate by using Azure Active Directory.
What should you do?

- A. Deploy the model to Azure Kubernetes Service (AKS). During deployment, set the token_auth_enabled parameter of the target configuration object to true
- B. Deploy the model to Azure Container Instances. During deployment, set the auth_enabled parameter of the target configuration object to true
- C. Deploy the model to Azure Container Instances. During deployment, set the token_auth_enabled parameter of the target configuration object to true
- D. Deploy the model to Azure Kubernetes Service (AKS). During deployment, set the auth.enabled parameter of the target configuration object to true

Correct Answer: *A*
To control token authentication, use the token_auth_enabled parameter when you create or update a deployment
Token authentication is disabled by default when you deploy to Azure Kubernetes Service.
Note: The model deployments created by Azure Machine Learning can be configured to use one of two authentication methods: key-based: A static key is used to authenticate to the web service.
token-based: A temporary token must be obtained from the Azure Machine Learning workspace (using Azure Active Directory) and used to authenticate to the web service.
Incorrect Answers:
C: Token authentication isn't supported when you deploy to Azure Container Instances.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-authenticate-web-service

Question #59*Topic 2*

HOTSPOT -
You are the owner of an Azure Machine Learning workspace.
You must prevent the creation or deletion of compute resources by using a custom role. You must allow all other operations inside the workspace.
You need to configure the custom role.
How should you complete the configuration? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

```
{
 "Name": "Data Scientist Custom",
 "IsCustom": true
 "Description": "Description"
 "Actions": [
          Microsoft.MachineLearningServices/workspaces/*/read
          Microsoft.MachineLearningServices/workspaces/computes/*/write
          Microsoft.MachineLearningServices/workspaces/delete

          Microsoft.MachineLearningServices/workspaces/*/write
          Microsoft.MachineLearningServices/workspaces/computes/*/write
          Microsoft.MachineLearningServices/workspaces/delete
 ],
 "NotActions": [
          Microsoft.MachineLearningServices/workspaces/*/read
          Microsoft.MachineLearningServices/workspaces/*/write
          Microsoft.MachineLearningServices/workspaces/computes/*/delete

          Microsoft.MachineLearningServices/workspaces/*/read
          Microsoft.MachineLearningServices/workspaces/*/write
          Microsoft.MachineLearningServices/workspaces/computes/*/write
 ],
 "AssignableScopes": [
   "/subscriptions/<subscription_id>"
 ]
}
```


Correct Answer:

Answer Area

```
{
 "Name": "Data Scientist Custom",
 "IsCustom": true
 "Description": "Description"
 "Actions": [
      Microsoft.MachineLearningServices/workspaces/*/read
      Microsoft.MachineLearningServices/workspaces/computes/*/write
      Microsoft.MachineLearningServices/workspaces/delete

      Microsoft.MachineLearningServices/workspaces/*/write
      Microsoft.MachineLearningServices/workspaces/computes/*/write
      Microsoft.MachineLearningServices/workspaces/delete
 ],
 "NotActions": [
      Microsoft.MachineLearningServices/workspaces/*/read
      Microsoft.MachineLearningServices/workspaces/*/write
      Microsoft.MachineLearningServices/workspaces/computes/*/delete

      Microsoft.MachineLearningServices/workspaces/*/read
      Microsoft.MachineLearningServices/workspaces/*/write
      Microsoft.MachineLearningServices/workspaces/computes/*/write
 ],
 "AssignableScopes": [
   "/subscriptions/<subscription_id>"
 ]
}
```

Box 1: Microsoft.MachineLearningServices/workspaces/*/read
Reader role: Read-only actions in the workspace. Readers can list and view assets, including datastore credentials, in a workspace. Readers can't create or update these assets.
Box 2: Microsoft.MachineLearningServices/workspaces/*/write
If the roles include Actions that have a wildcard (*), the effective permissions are computed by subtracting the NotActions from the allowed Actions.
Box 3: Box 2: Microsoft.MachineLearningServices/workspaces/computes/*/delete
Box 4: Microsoft.MachineLearningServices/workspaces/computes/*/write
Reference:
https://docs.microsoft.com/en-us/azure/role-based-access-control/overview#how-azure-rbac-determines-if-a-user-has-access-to-a-resource
Question #60*Topic 2*

HOTSPOT -
You create an Azure Machine Learning workspace named workspace1. You assign a custom role to a user of workspace1.
The custom role has the following JSON definition:

```
{
 "Name": "MyRole",
 "IsCustom": true,
 "Description": "New custom role description.",
 "Actions": ["*"],
 "NotActions": [
   "Microsoft.MachineLearningServices/workspaces/write",
   "Microsoft.MachineLearningServices/workspaces/computes/*/write",
   "Microsoft.MachineLearningServices/workspaces/computes/*/delete",
   "Microsoft. Authorization/*/write"
 ],
 "AssignableScopes": [
   "/subscriptions/<subscription_id>/resourceGroups/resourcegroup1/providers/
   Microsoft.MachineLearningServices/workspaces/workspace1"
 ]
}
```

Instructions: For each of the following statements, select Yes if the statement is true. Otherwise, select No.
NOTE: Each correct selection is worth one point.
Hot Area:

Answer Area

Statements	Yes	No
The user can perform all actions in the workspace		
The user can delete a compute resource in the workspace		
The user can write metrics to the workspace		

Correct Answer:

Answer Area

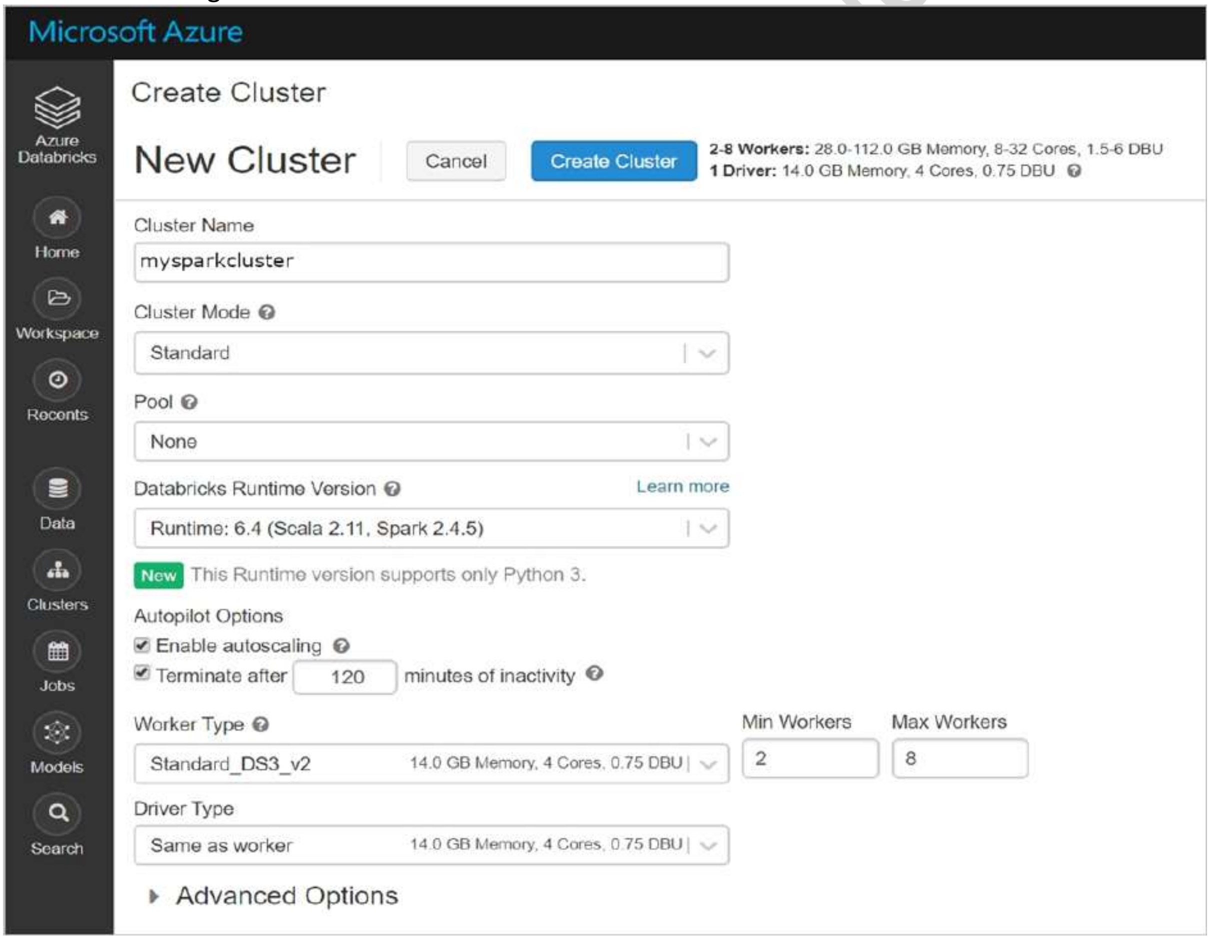
Statements	Yes	No
The user can perform all actions in the workspace		
The user can delete a compute resource in the workspace		
The user can write metrics to the workspace		

Box 1: No -
The actions listed in NotActions are prohibited.
If the roles include Actions that have a wildcard (*), the effective permissions are computed by subtracting the NotActions from the allowed Actions.

Box 2: No -
Deleting compute resources in the workspace is in the NotActions list.

Box 3: Yes -
Writing metrics is not listed in NotActions.
Reference:
https://docs.microsoft.com/en-us/azure/role-based-access-control/overview#how-azure-rbac-determines-if-a-user-has-access-to-a-resource
Question #61*Topic 2*

HOTSPOT -
You create a new Azure Databricks workspace.
You configure a new cluster for long-running tasks with mixed loads on the compute cluster as shown in the image below.



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

Hot Area:

Answer Area

Code for each user runs as a separate process	Yes / No
The number of workers is fixed for the entire duration of the job	Yes / No

Correct Answer:

Answer Area

Code for each user runs as a separate process	Yes / **No**
The number of workers is fixed for the entire duration of the job	Yes / **No**

Box 1: No -
Running user code in separate processes is not possible in Scala.

Box 2: No -
Autoscaling is enabled. Minimum 2 workers, Maximum 8 workers.
Reference:
https://docs.databricks.com/clusters/configure.html
Question #62*Topic 2*

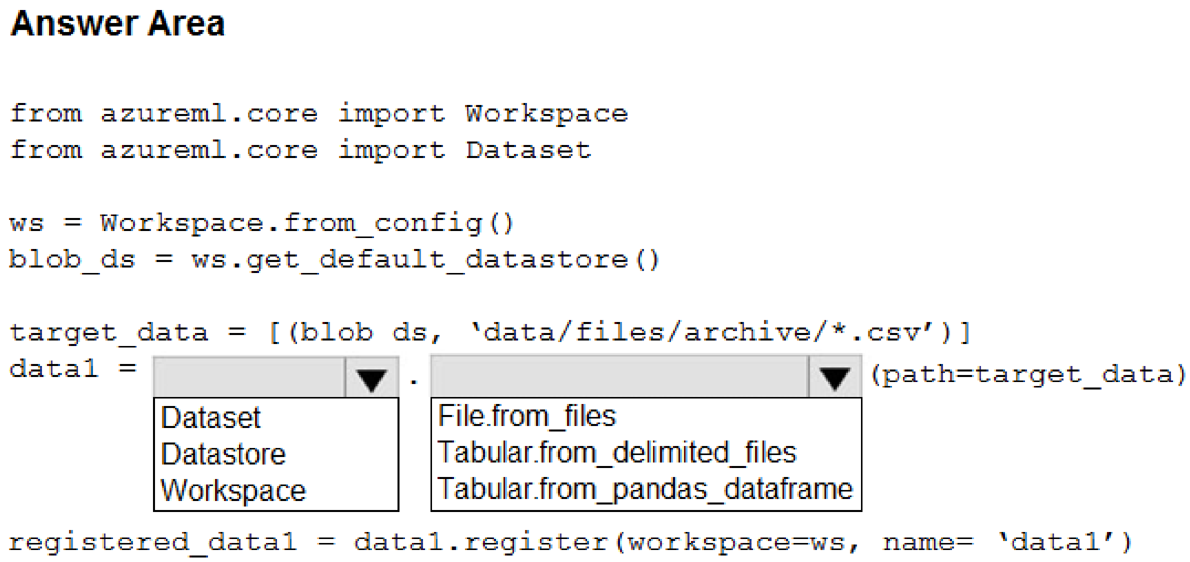
HOTSPOT
-

You use an Azure Machine Learning workspace. The default datastore contains comma-separated values (CSV) files.

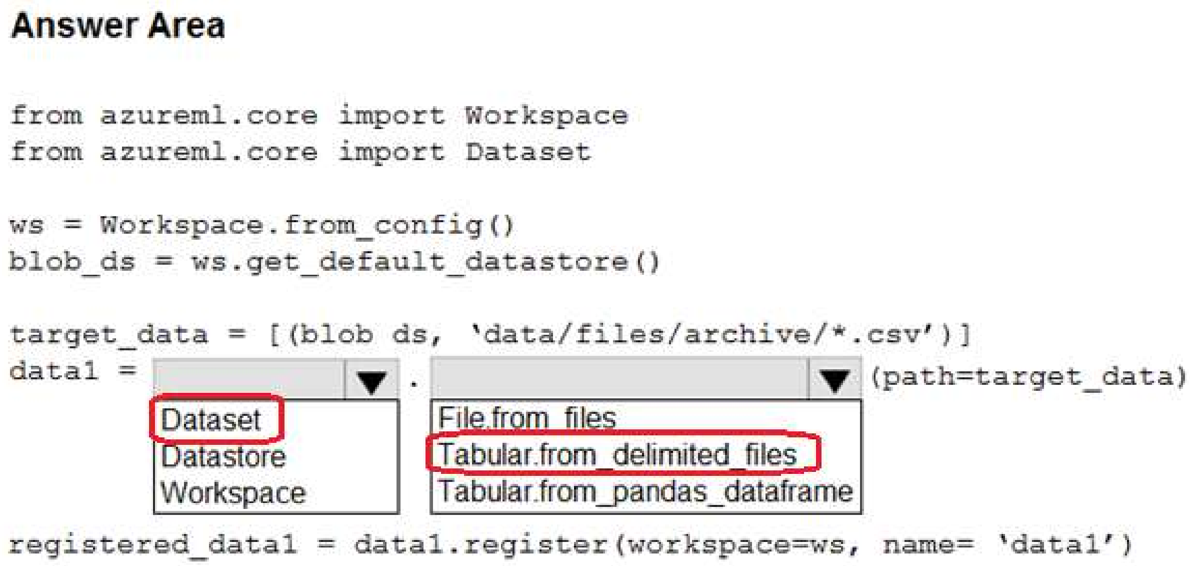
The CSV files must be made available for use in experiments and data processing pipelines. The files must be loaded directly into pandas dataframes.

How should you complete the code? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.



Correct Answer:



Question #63*Topic 2*

HOTSPOT

-

You plan to use a curated environment to run Azure Machine Learning training experiments in a workspace.

You need to display all curated environments and their respective packages in the workspace.

How should you complete the code? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Answer Area

```
from azureml.core import Workspace, Environment
ws = Workspace.from_config()
envs = Environment.list(workspace=ws)
for env in envs:

    if env.startswith("  [ AutoML | AzureML | PyTorch ]  "):

        print("Name", env)
        print("packages", envs[env].  [ python | version | docker ]  .conda_dependencies.serialize_to_string())
```

Correct Answer:

Answer Area

```
from azureml.core import Workspace, Environment
ws = Workspace.from_config()
envs = Environment.list(workspace=ws)
for env in envs:

    if env.startswith("  [ AutoML | AzureML | PyTorch ]  "):

        print("Name", env)
        print("packages", envs[env].  [ python | version | docker ]  .conda_dependencies.serialize_to_string())
```

Question #64*Topic 2*

You are profiling data by using Azure Machine Learning studio.

You need to detect columns with odd or missing values.

Which statistic should you analyze?

- A. Profile
- B. Std deviation
- C. Error count
- D. Type

Correct Answer: *C*

Question #65*Topic 2*

You are authoring a notebook in Azure Machine Learning studio.

You must install packages from the notebook into the currently running kernel. The installation must be limited to the currently running kernel only.

You need to install the packages.

Which magic function should you use?

- A. !pip
- B. %pip
- C. !conda
- D. %load

Correct Answer: *B*

Question #66*Topic 2*

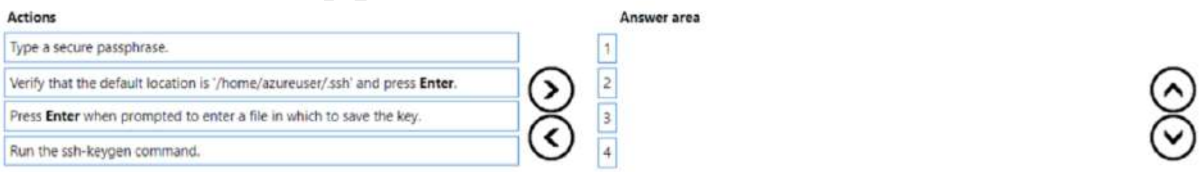
DRAG DROP
-

You need to implement source control for scripts in an Azure Machine Learning workspace. You use a terminal window in the Azure Machine Learning Notebook tab.

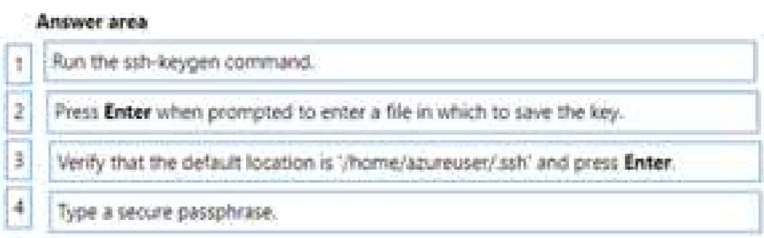
You must authenticate your Git account with SSH.

You need to generate a new SSH key.

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.



Correct Answer:



Question #67*Topic 2*

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 use Azure Machine Learning designer to load the following datasets into an experiment:

Dataset1 -

Age	Length	Width
3	22	13
7	11	96
18	32	85

Dataset2 -

Age	Length	Width
11	101	65
6	98	23
33	22	54
17	52	12

You need to create a dataset that has the same columns and header row as the input datasets and contains all rows from both input datasets.

Solution: Use the Add Rows module.

Does the solution meet the goal?

- A. Yes
- B. No

Correct Answer: *A*

Question #68*Topic 2*

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 use Azure Machine Learning designer to load the following datasets into an experiment:

Dataset1 -

Age	Length	Width
3	22	13
7	11	96
18	32	85

Dataset2 -

Age	Length	Width
11	101	65
6	98	23
33	22	54
17	52	12

You need to create a dataset that has the same columns and header row as the input datasets and contains all rows from both input datasets.

Solution: Use the Apply Transformation module.

Does the solution meet the goal?

- A. Yes
- B. No

Correct Answer: *B*

Question #69*Topic 2*

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 use Azure Machine Learning designer to load the following datasets into an experiment:

Dataset1 -

Age	Length	Width
3	22	13
7	11	96
18	32	85

Dataset2 -

Age	Length	Width
11	101	65
6	98	23
33	22	54
17	52	12

You need to create a dataset that has the same columns and header row as the input datasets and contains all rows from both input datasets.

Solution: Use the Execute Python Script module.

Does the solution meet the goal?

- A. Yes
- B. No

Correct Answer: *A*

Question #70*Topic 2*

HOTSPOT
-

You must use an Azure Data Science Virtual Machine (DSVM) as a compute target.

You need to attach an existing DSVM to the workspace by using the Azure Machine Learning SDK for Python.

How should you complete the following code segment? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Answer Area

```
from azureml.core.compute import RemoteCompute, ComputeTarget
compute_target_name = "dsvm"
config = RemoteCompute.[ ▼ ](resource_id='<resource_id>',
```

attach_configuration
get_credentials
detach

```
 ssh_port=22, username='<username>', private_key_file='./.ssh/id_rsa')
compute = ComputeTarget.[ ▼ ] (ws, compute_target_name, config)
```

detach
create
attach

```
compute.wait_for_completion(show_output=True)
```

Correct Answer:

Answer Area

```
from azureml.core.compute import RemoteCompute, ComputeTarget
compute_target_name = "dsvm"
config = RemoteCompute.[ ▼ ](resource_id='<resource_id>',
```

attach_configuration
get_credentials
detach

```
 ssh_port=22, username='<username>', private_key_file='./.ssh/id_rsa')
compute = ComputeTarget.[ ▼ ] (ws, compute_target_name, config)
```

detach
create
attach

```
compute.wait_for_completion(show_output=True)
```

Question #71*Topic 2*

HOTSPOT

-

You have an Azure Machine Learning workspace.

You run the following code in a Python environment in which the configuration file for your workspace has been downloaded.

```
from azureml.core import Workspace
from azureml.core import Experiment
import pandas as pd
import datetime as dt
ws = Workspace.from_config()
experiment = Experiment(workspace=ws, name= 'my_experiment')
run = experiment.start_logging()
print('run_time', dt.datetime.now())

row_count = (len(data))
run.log('observations', row_count)
run.complete()
```

Instructions: For each of the following statements, select Yes if the statement is true. Otherwise, select No.

NOTE: Each correct selection is worth one point.

Answer Area

Statements	Yes	No
An error will occur if an experiment named my_experiment does not already exist in the workspace.	○	○
If the experiment does not exist, it will be created. If the experiment does exist, the code will create a new run of the existing experiment.	○	○
After the code completes, a metric named run_time is recorded in the experiment run. The metric will contain the date and time for the run.	○	○
After the code completes, the data.csv file will be available in the run's output.	○	○

Correct Answer:

Answer Area

Statements	Yes	No
An error will occur if an experiment named my_experiment does not already exist in the workspace.	○	$\boxed{\circ}$
If the experiment does not exist, it will be created. If the experiment does exist, the code will create a new run of the existing experiment.	$\boxed{\circ}$	○
After the code completes, a metric named run_time is recorded in the experiment run. The metric will contain the date and time for the run.	○	$\boxed{\circ}$
After the code completes, the data.csv file will be available in the run's output.	○	$\boxed{\circ}$

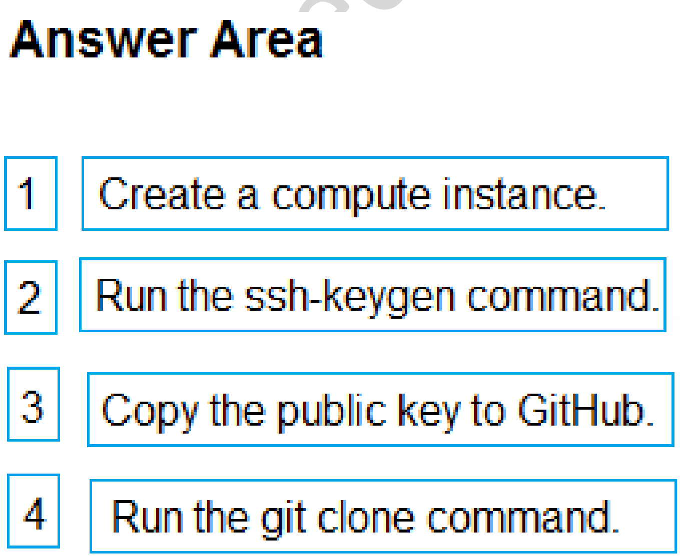
Question #72*Topic 2*

DRAG DROP
-

You create an Azure Machine Learning workspace.

You need to use the shared file system of the workspace to store a clone of a private Git repository.

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.



Correct Answer:

Question #73*Topic 2*

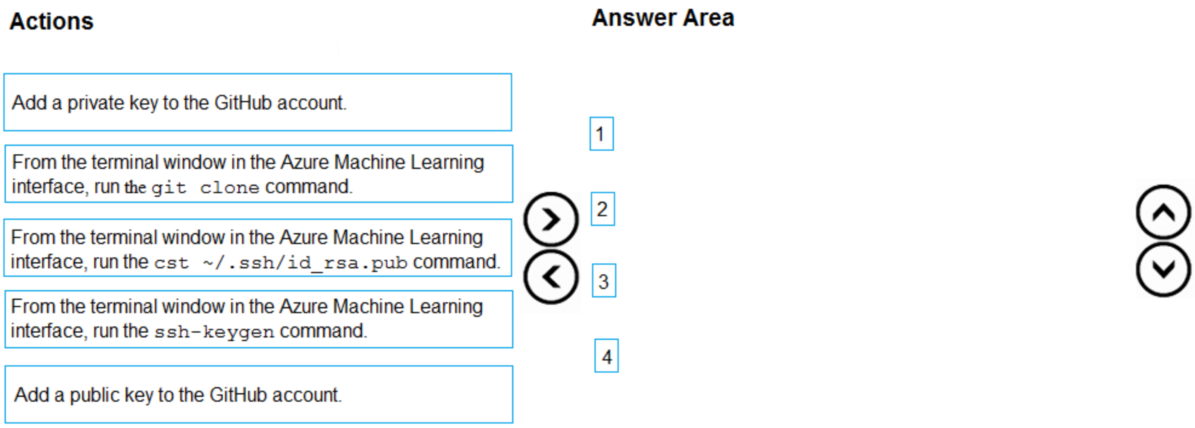
DRAG DROP
-

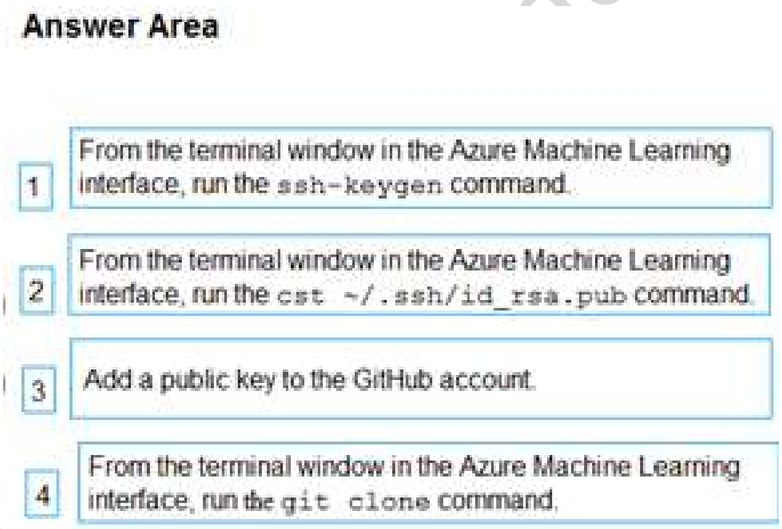
You have an existing GitHub repository containing Azure Machine Learning project files.

You need to clone the repository to your Azure Machine Learning shared workspace file system.

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.

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





Correct Answer:

Question #74*Topic 2*

You are creating a compute target to train a machine learning experiment.

The compute target must support automated machine learning, machine learning pipelines, and Azure Machine Learning designer training.

You need to configure the compute target.

Which option should you use?

- A. Azure HDInsight
- B. Azure Machine Learning compute cluster
- C. Azure Batch

- D. Remote VM

Correct Answer: *B*

Question #75*Topic 2*

You manage an Azure Machine Learning workspace by using the Azure CLI ml extension v2.

You need to define a YAML schema to create a compute cluster.

Which schema should you use?

- A. https://azuremlschemas.azureedge.net/latest/computeInstance.schema.json
- B. https://azuremlschemas.azureedge.net/latest/amlCompute.schema.json
- C. https://azuremlschemas.azureedge.net/latest/vmCompute.schema.json
- D. https://azuremlschemas.azureedge.net/latest/kubernetesCompute.schema.json

Correct Answer: *B*

Question #76*Topic 2*

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 have the following Azure subscriptions and Azure Machine Learning service workspaces:

Subscription	Workspace	Comment
385bdfe5-4cef-4ad4-b977-3f86d92727c9	ml-default	This is default subscription.
5a5891d1-557a-4234-9b83-2e90412b1068	ml-project	The information required to uniquely identify this workspace is stored in the file config.json in the same folder as the Python script.

You need to obtain a reference to the ml-project workspace.

Solution: Run the following Python code:

```
from azureml.core import Workspace
ws = Workspace.from_config()
```

Does the solution meet the goal?

- A. Yes
- B. No

Correct Answer: *A*

Question #77*Topic 2*

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 have the following Azure subscriptions and Azure Machine Learning service workspaces:

Subscription	Workspace	Comment
385bdfe5-4cef-4ad4-b977-3f86d92727c9	ml-default	This is default subscription.
5a5891d1-557a-4234-9b83-2e90412b1068	ml-project	The information required to uniquely identify this workspace is stored in the file config.json in the same folder as the Python script.

You need to obtain a reference to the ml-project workspace.

Solution: Run the following Python code:

```
from azureml.core import Workspace
ws = Workspace.get(name="ml-project")
```

Does the solution meet the goal?

- A. Yes
- B. No

Correct Answer: *B*

Question #78*Topic 2*

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 have the following Azure subscriptions and Azure Machine Learning service workspaces:

Subscription	Workspace	Comment
385bdfe5-4cef-4ad4-b977-3f86d92727c9	ml-default	This is default subscription.
5a5891d1-557a-4234-9b83-2e90412b1068	ml-project	The information required to uniquely identify this workspace is stored in the file config.json in the same folder as the Python script.

You need to obtain a reference to the ml-project workspace.

Solution: Run the following Python code:

```
from azureml.core import Workspace
ws = Workspace(workspace_name="ml-project")
```

Does the solution meet the goal?

- A. Yes
- B. No

Correct Answer: *B*

Question #79*Topic 2*

HOTSPOT

-

You create an Azure Machine Learning workspace. You use the Azure Machine Learning SDK for Python.

You must create a dataset from remote paths. The dataset must be reusable within the workspace.

You need to create the dataset.

How should you complete the following code segment? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Answer Area

```
from azureml.core import Dataset
from azureml.data.dataset_factory import DataType
web_paths = ['https://domain.blob.core.windows.net/demo/dataset1.tsv',
   'https://domain.blob.core.windows.net/demo/dataset2.tsv']

ds = Dataset [ ▼ ] (path=web_paths)
```

Tabular.from_delimited_files
Tabular.from_parquet_files

```
ds = ds. [ ▼ ] (workspace=workspace,
```

update
register
unregister_all_versions

```
     name= 'ds',
     description= 'training data')
```

Correct Answer:

Answer Area

```
from azureml.core import Dataset
from azureml.data.dataset_factory import DataType
web_paths = ['https://domain.blob.core.windows.net/demo/dataset1.tsv',
   'https://domain.blob.core.windows.net/demo/dataset2.tsv']

ds = Dataset [ ▼ ] (path=web_paths)
```

Tabular.from_delimited_files
Tabular.from_parquet_files

```
ds = ds. [ ▼ ] (workspace=workspace,
```

update
register
unregister_all_versions

```
     name= 'ds',
     description= 'training data')
```

Question #80*Topic 2*

HOTSPOT
-

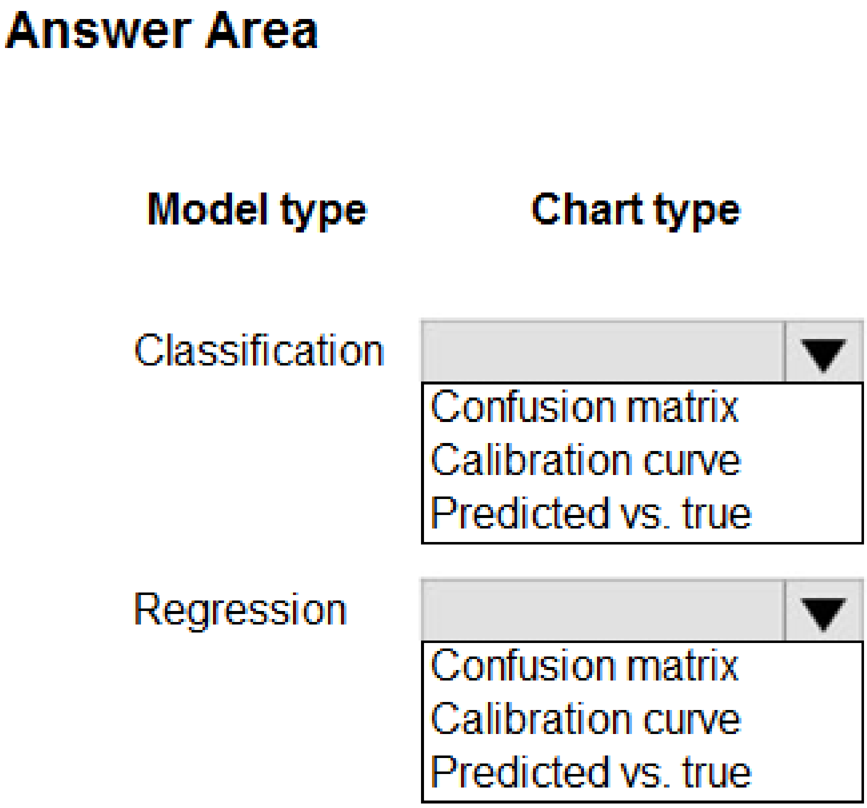
You train classification and regression models by using automated machine learning.

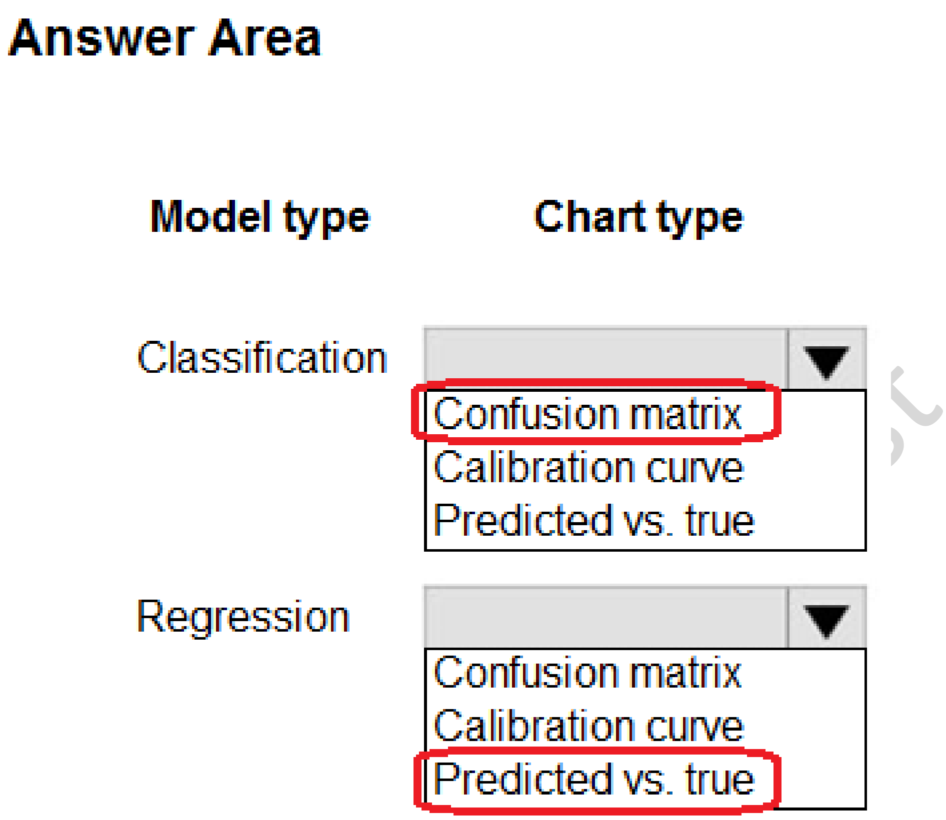
You must evaluate automated machine learning experiment results. The results include how a classification model is making systematic errors in its predictions and the relationship between the target feature and the regression model's predictions. You must use charts generated by automated machine learning.

You need to choose a chart type for each model type.

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

NOTE: Each correct selection is worth one point.





Correct Answer:

Question #81*Topic 2*

HOTSPOT

-

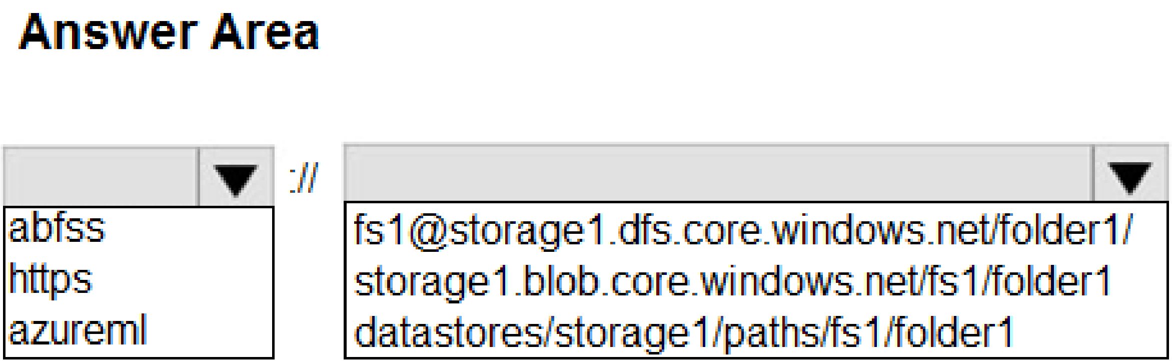
You create an Azure Data Lake Storage Gen2 storage account named storage1containing a file system named fs1 and a folder named folder1.

The contents of folder1 must be accessible from jobs on compute targets in the Azure Machine Learning workspace.

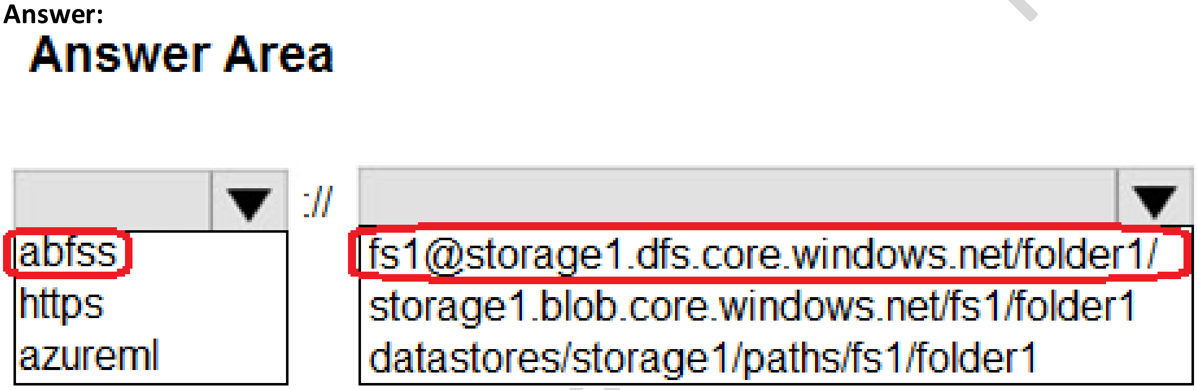
You need to construct a URI to reference folder1.

How should you construct the URI? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.



Correct Answer:



Question #82*Topic 2*

HOTSPOT

-

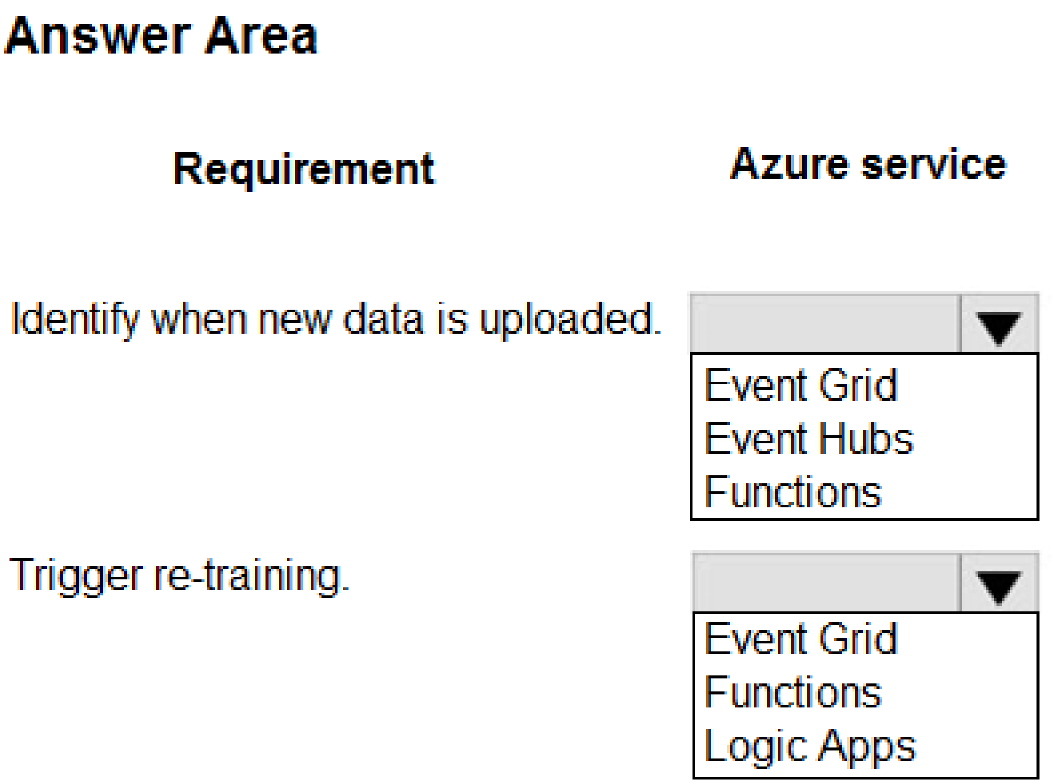
You train a model by using Azure Machine Learning. You use Azure Blob Storage to store production data.

The model must be re-trained when new data is uploaded to Azure Blob Storage. You need to minimize development and coding.

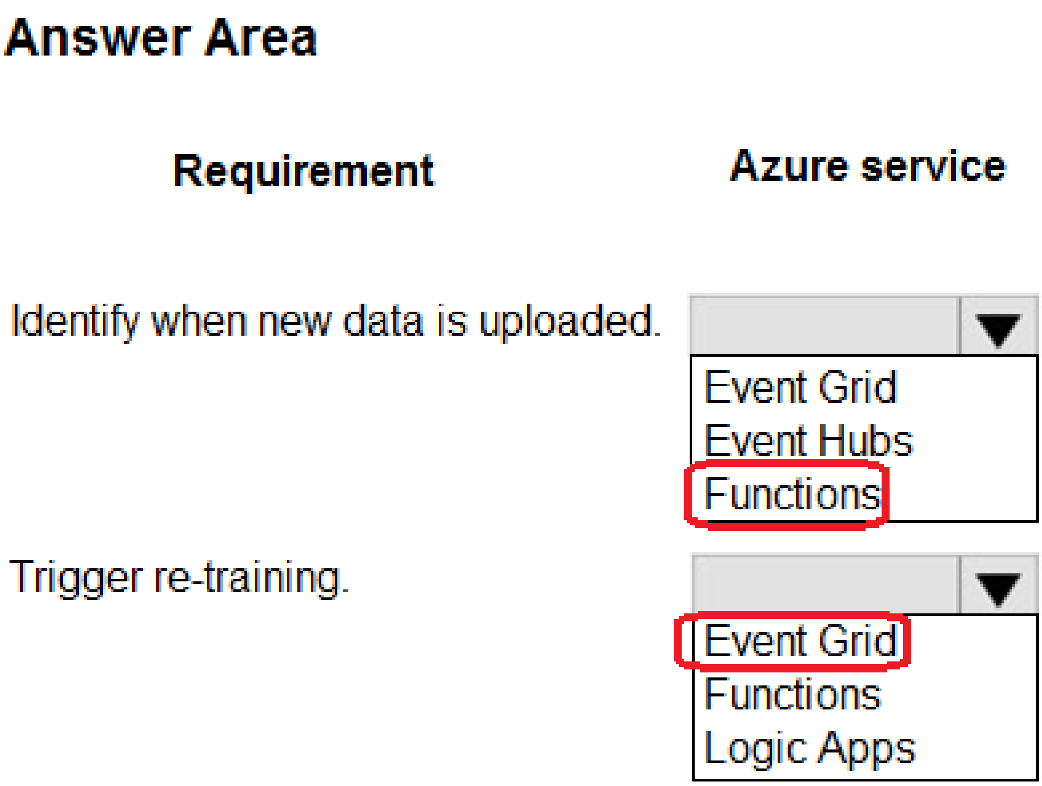
You need to configure Azure services to develop a re-training solution.

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

NOTE: Each correct selection is worth one point.



Correct



Answer:

Question #83*Topic 2*

You use the Azure Machine Learning SDK for Python v1 and notebooks to train a model. You create a compute target, an environment, and a training script by using Python code.

You need to prepare information to submit a training run.

Which class should you use?

- A. ScriptRun
- B. ScriptRunConfig
- C. RunConfiguration
- D. Run

Correct Answer: *B*

Question #84*Topic 2*

HOTSPOT
-

You have an Azure Machine Learning workspace.

You need to use the Azure Machine Learning SDK for Python to create and register the Azure Data Lake Storage Generation 2 datastore for the workspace.

How should you complete the following code segment? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Answer Area

```
import azureml.core
from azureml.core import Workspace, Datastore
ws = Workspace.from_config()

datastore = Datastore.[dropdown] (
```

- register_azure_data_lake_gen2
- register_dbfs
- register_azure_blob_container

```
    workspace=ws,
    datastore_name='datastore_name',
    account_name='account_name',
    [dropdown] ,
```

- database_name=database_name
- filesystem='test'
- container_name=container_name

```
    tenant_id=tenant_id,
    client_id=client_id,
    client_secret=client_secret)
```


Correct Answer:

Answer Area

```
import azureml.core
from azureml.core import Workspace, Datastore
ws = Workspace.from_config()

datastore = Datastore.[                    ▼] (
                     register_azure_data_lake_gen2
                     register_dbfs
                     register_azure_blob_container

    workspace=ws,
    datastore_name='datastore_name',
    account_name='account_name',
    [                    ▼] ,
    database_name=database_name
    filesystem='test'
    container_name=container_name
    tenant_id=tenant_id,
    client_id=client_id,
    client_secret=client_secret)
```


Question #85*Topic 2*

HOTSPOT

-

You manage an Azure Machine Learning workspace. You create a training script named sample_training_script.py. The script is used to train a predictive model in the conda environment defined by a file named environment.yml.

You need to run the script as an experiment.

How should you complete the following code segment? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Answer Area

```
import azureml.core
from azureml.core import Workspace
from azureml.core import Experiment, ScriptRunConfig, Environment
from azureml.widgets import RunDetails
ws = Workspace.from_config()
env = Environment.from_conda_specification("experiment_env", "environment.yml")
object1 = [ ▼ ] (source_directory='sample_folder',
          ScriptRun         script='sample_training_script.py',
          ScriptRunConfig   environment=env)
          RunConfiguration

object2 = Experiment(workspace=ws, name='sample_object2')
run = object2.[ ▼ ] (object1)
             get_runs
             reactivate
             submit

run.wait_for_completion()
```

Correct Answer:

Answer Area

```
import azureml.core
from azureml.core import Workspace
from azureml.core import Experiment, ScriptRunConfig, Environment
from azureml.widgets import RunDetails
ws = Workspace.from_config()
env = Environment.from_conda_specification("experiment_env", "environment.yml")
object1 = [ ▼ ] (source_directory='sample_folder',
          ScriptRun         script='sample_training_script.py',
          ScriptRunConfig   environment=env)
          RunConfiguration

object2 = Experiment(workspace=ws, name='sample_object2')
run = object2.[ ▼ ] (object1)
             get_runs
             reactivate
             submit

run.wait_for_completion()
```

Question #86*Topic 2*

You have an Azure Machine Learning workspace. You are connecting an Azure Data Lake Storage Gen2 account to the workspace as a data store.

You need to authorize access from the workspace to the Azure Data Lake Storage Gen2 account.

What should you use?

- A. Service principal
- B. SAS token
- C. Managed identity
- D. Account key

Correct Answer: *A*

Question #87*Topic 2*

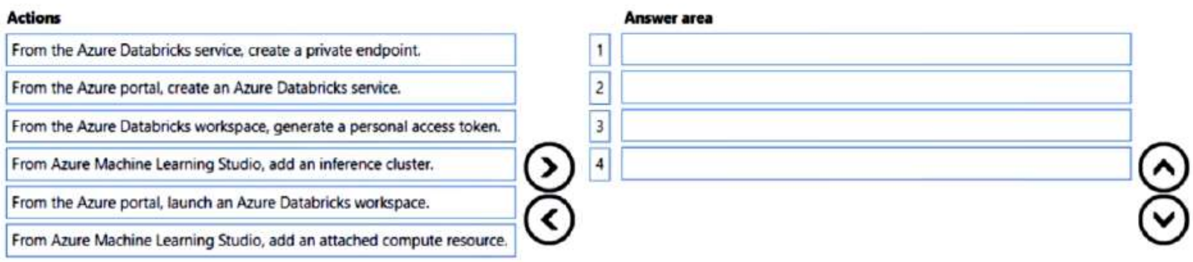
DRAG DROP

-

You provision an Azure Machine Learning workspace in a new Azure subscription.

You need to attach Azure Databricks as a compute resource from the Azure Machine Learning workspace.

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.



Correct Answer:

Question #88*Topic 2*

HOTSPOT

-

You are designing a machine learning solution.

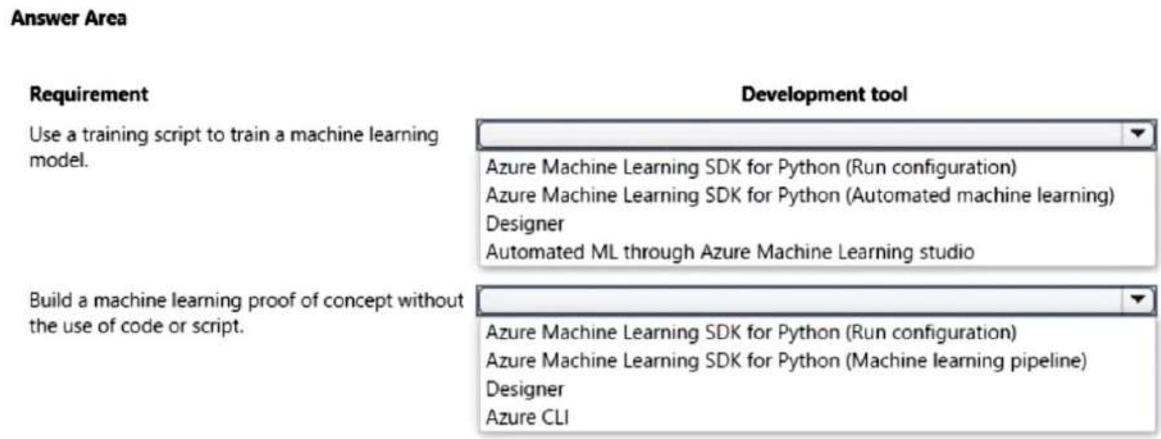
You have the following requirements:

• Use a training script to train a machine learning model.
• Build a machine learning proof of concept without the use of code or script.

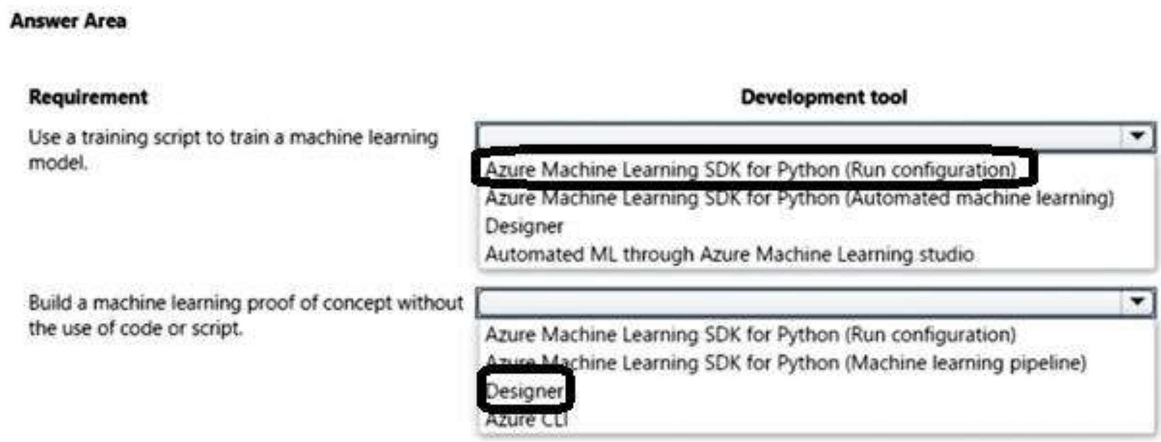
You need to select a development tool for each requirement.

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

NOTE: Each correct selection is worth one point.



Correct Answer:



Question #89 *Topic 2*

You manage an Azure Machine Learning workspace named workspace1.

You must develop Python SDK v2 code to attach an Azure Synapse Spark pool as a compute target in workspace1. The code must invoke the constructor of the SynapseSparkCompute class.

You need to invoke the constructor.

What should you use?

- A. Synapse workspace web URL and Spark pool name
- B. resource ID of the Synapse Spark pool and a user-defined name
- C. pool URL of the Synapse Spark pool and a system-assigned name
- D. Synapse workspace name and workspace web URL

Correct Answer: *B*

Question #90*Topic 2*

HOTSPOT
-

You manage an Azure Machine Learning workspace by using the Python SDK v2.

You must create a compute cluster in the workspace. The compute cluster must run workloads and property handle interruptions. You start by calculating the maximum amount of compute resources required by the workloads and size the cluster to match the calculations.

The cluster definition includes the following properties and values:
• names="mlcluster"
• size="STANDARD_DS3_v2"
• min_instances=1
• max_instances=4
• tier="dedicated"

The cost of the compute resources must be minimized when a workload is active or idle. Cluster property changes must not affect the maximum amount of compute resources available to the workloads run on the cluster.

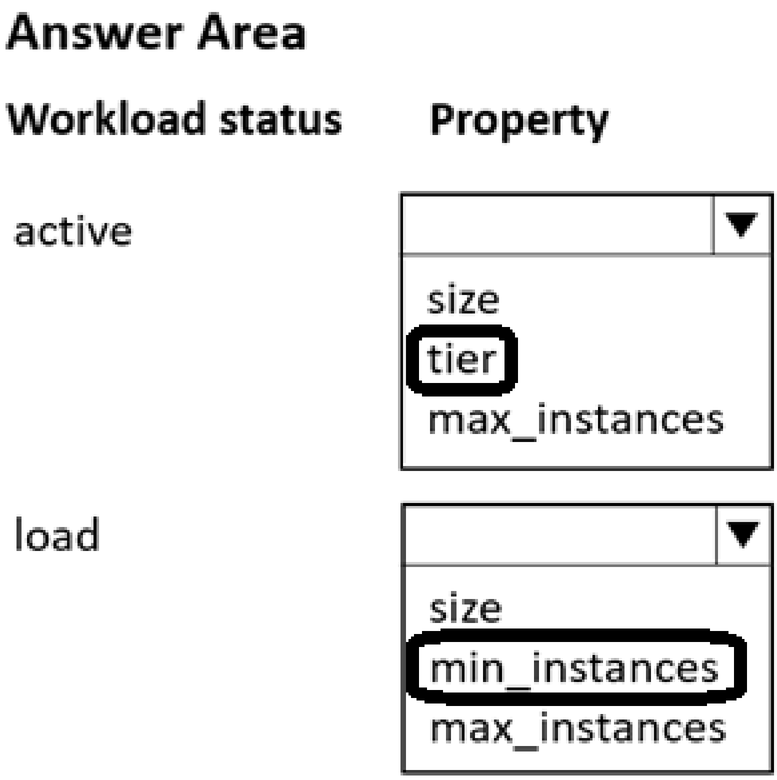
You need to modify the cluster properties to minimize the cost of compute resources.

Which properties should you modify? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Answer Area

Workload status	Property
active	size tier max_instances
load	size min_instances max_instances



Correct Answer:

Question #91*Topic 2*

HOTSPOT

-

You manage an Azure Machine Learning workspace. You create an experiment named experiment by using the Azure Machine Learning Python SDK v2 and MLflow.

You are reviewing the results of experiment by using the following code segment:

```
runs = mlflow.search_runs(
    experiment_names=["experiment1"],
    max_results=5,
    order_by=["start_time ASC"])

runs[runs.status == "FAILED"]
```

For each of the following statements, select Yes if the statement is true. Otherwise, select No.

Answer Area

Statements	Yes	No
Aborted runs are returned.	○	○
The latest five experiment runs are returned.	○	○
The jobs that are returned have been canceled or killed by the user or system.	○	○
All metrics and their values are returned for the returned experiment runs.	○	○

Correct Answer:

Answer Area

Statements	Yes	No
Aborted runs are returned.	○	○
The latest five experiment runs are returned.	○	○
The jobs that are returned have been canceled or killed by the user or system.	○	○
All metrics and their values are returned for the returned experiment runs.	○	○

Question #92*Topic 2*

You manage an Azure Machine Learning workspace. You have an environment for training jobs which uses an existing Docker image.

A new version of the Docker image is available.

You need to use the latest version of the Docker image for the environment configuration by using the Azure Machine Learning SDK v2.

What should you do?

- A. Modify the conda_file to specify the new version of the Docker image.
- B. Use the Environment class to create a new version of the environment.
- C. Use the create_or_update method to change the tag of the image.
- D. Change the description parameter of the environment configuration.

Correct Answer: *B*

Question #93*Topic 2*

HOTSPOT

-

You manage an Azure Machine Learning workspace by using the Python SDK v2.

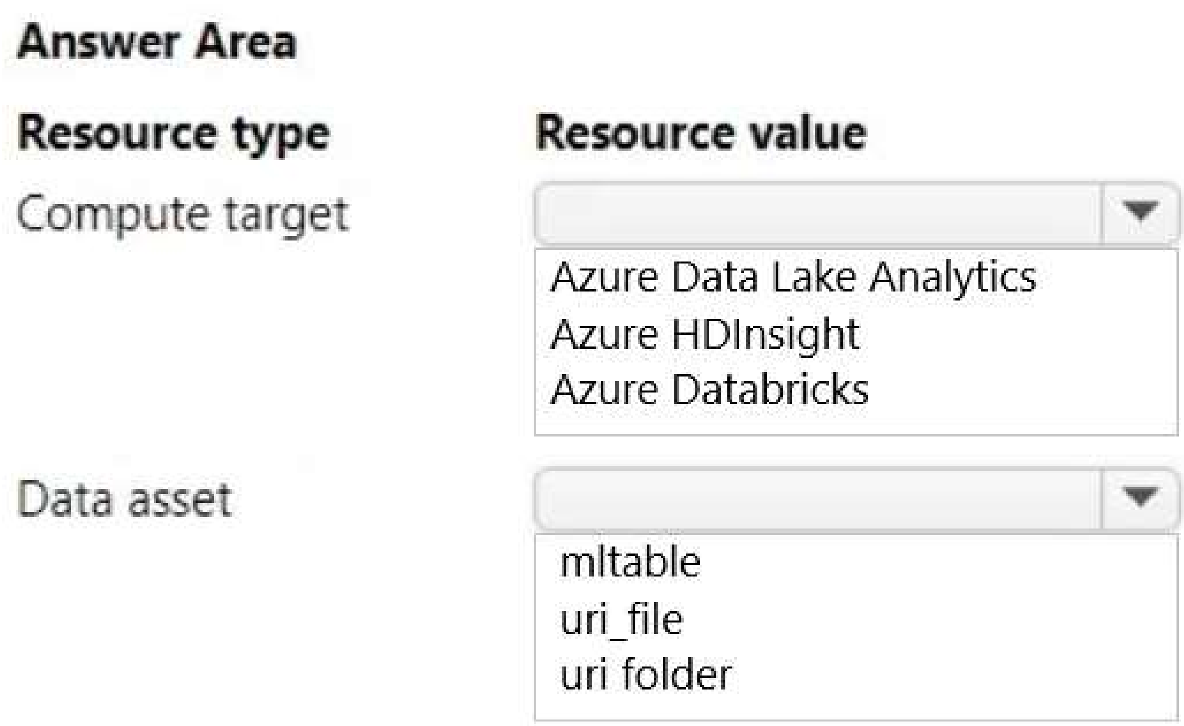
You must create an automated machine learning job to generate a classification model by using data files stored in Parquet format.

You must configure an autoscaling compute target and a data asset for the job.

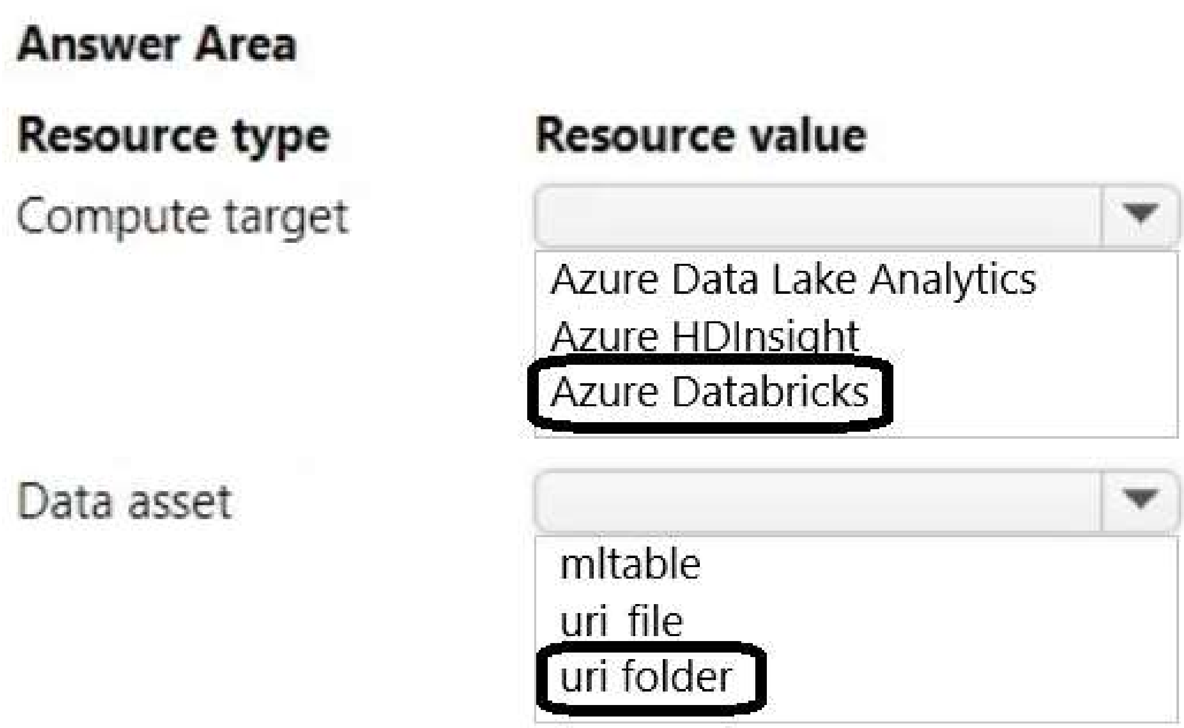
You need to configure the resources for the job.

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

NOTE: Each correct selection is worth one point.



Correct Answer:



Question #94*Topic 2*

HOTSPOT

-

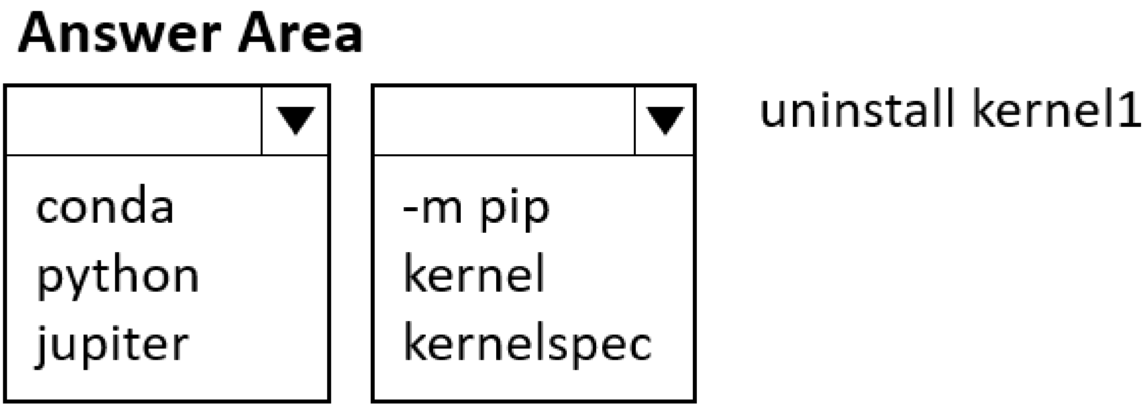
You manage an Azure Machine Learning workspace named workspace1 with a compute instance named compute1.

You must remove a kernel named kernel1 from compute1. You connect to compute1 by using a terminal window from workspace1.

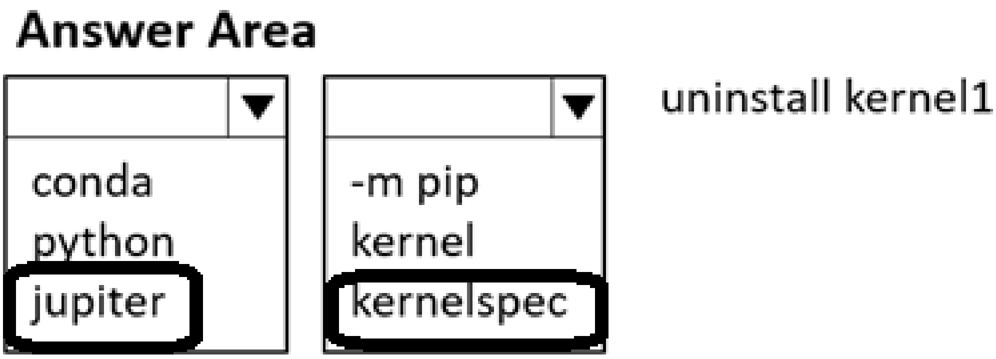
You need to enter a command in the terminal window to remove kernel.

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

NOTE: Each correct selection is worth one point.



Correct



Answer:

Question #95*Topic 2*

You manage an Azure Machine Learning workspace. The workspace includes an Azure Machine Learning Kubernetes compute target configured as an Azure Kubernetes Service (AKS) cluster named AKS1. AKS1 is configured to enable the targeting of different nodes to train workloads.

You must run a command job on AKS1 by using the Azure ML Python SDK v2. The command job must select different types of compute nodes. The compute node types must be specified by using a command parameter.

You need to configure the command parameter.

Which parameter should you use?

- A. environment
- B. compute
- C. limits
- D. instance_type

Correct Answer: *D*

Question #96*Topic 2*

HOTSPOT

-

You create a new Azure Machine Learning workspace with a compute cluster.

You need to create the compute cluster asynchronously by using the Azure Machine Learning Python SDK v2.

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

cluster = [▼] (

- ComputeConfiguration
- AmlCompute
- ComputeInstance

```
    name=cpu_cluster_name,
    size="STANDARD_DS3_v2",
    max_instances=4
)
```

[▼] (cluster)

- ml_client.from_config
- ml_client.create_or_update
- ml_client.begin_create_or_update

Correct Answer:

Answer Area

cluster = [▼] (

- ComputeConfiguration
- AmlCompute
- ComputeInstance

name=cpu_cluster_name,
size="STANDARD_DS3_v2",
max_instances=4
)

[▼] (cluster)

- ml_client.from_config
- ml_client.create_or_update
- ml_client.begin_create_or_update

Question #97*Topic 2*

DRAG DROP

-

You manage an Azure Machine Learning workspace named workspace1 by using the Python SDK v2.

You must register datastores in workspace1 for Azure Blob storage and Azure Files storage to meet the following requirements:

- Azure Active Directory (Azure AD) authentication must be used for access to storage when possible.
- Credentials and secrets stored in workspace1 must be valid for a specified time period when accessing Azure Files storage.

You need to configure a security access method used to register the Azure Blob and Azure Files storage in workspace1.

Which security access method should you configure? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Security access methods

- Account key
- Identity-based access
- Shared Access Signature (SAS)

Answer Area

Storage type	**Security access method**
Azure Blob storage	
Azure Files storage	

Correct

Answer Area

Storage type	**Security access method**
Azure Blob storage	Identity-based access
Azure Files storage	Shared Access Signature (SAS)

Answer:

Question #98*Topic 2*

You manage an Azure Machine Learning workspace named workspace1.

You must develop Python SDK v2 code to add a compute instance to workspace1. The code must import all required modules and call the constructor of the ComputeInstance class.

You need to add the instantiated compute instance to workspace1.

What should you use?

- A. constructor of the azure.ai.ml.ComputeSchedule class
- B. constructor of the azure.ai.ml.ComputePowerAction enum
- C. begin_create_or_update method of an instance of the azure.ai.ml.MLClient class
- D. set_resources method of an instance of the azure.ai.ml.Command class

Correct Answer: *C*

Question #99*Topic 2*

HOTSPOT

-

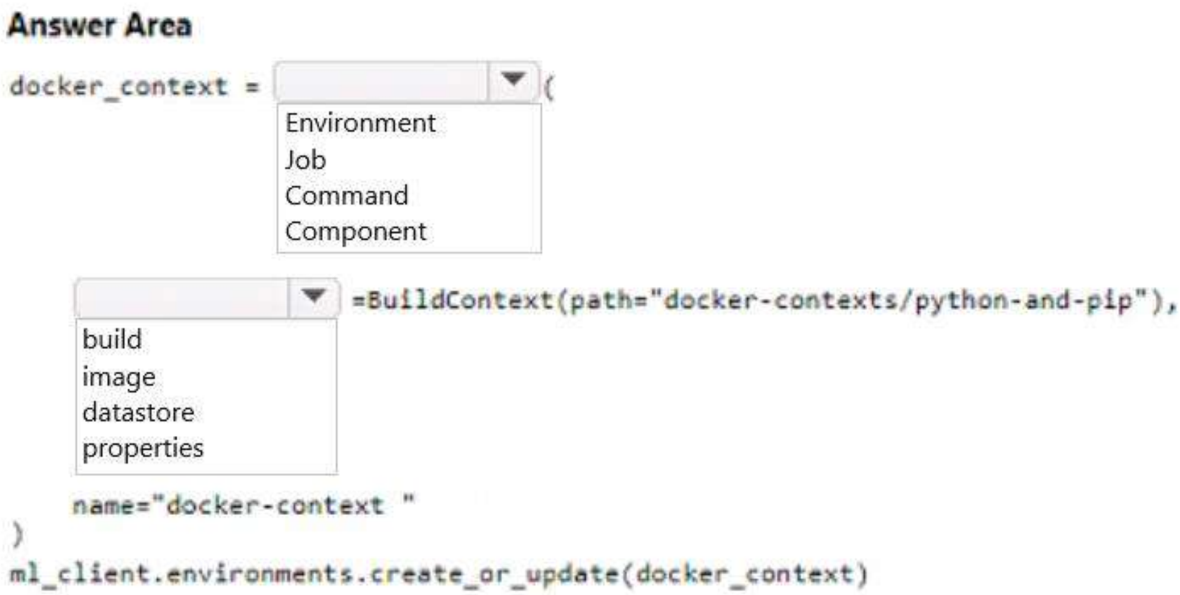
You manage an Azure Machine Learning workspace.

You must define the execution environments for your jobs and encapsulate the dependencies for your code.

You need to configure the environment from a Docker build context.

How should you complete the code segment? To answer, select the appropriate option in the answer area.

NOTE: Each correct selection is worth one point.



Correct Answer:

Question #100*Topic 2*

HOTSPOT
-

You create an Azure Machine Learning workspace. You use the Azure Machine Learning Python SDK v2 to create a compute cluster.

The compute cluster must run a training script. Costs associated with running the training script must be minimized.

You need to complete the Python script to create the compute cluster.

How should you complete the script? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Answer Area

```
from azure.ai.ml.entities import AmlCompute
try:
    cpu_cluster = ml_client.compute.get("cpu-cluster")
except Exception:
    cpu_cluster = [ ▼ ] (
```

- AmlCompute
- ComputeInstance
- KubernetesCompute

```
        name="cpu-cluster",
        size="STANDARD_DS3_V2",
        max_instances=4,
        [ ▼ ] ,
```

- tier="LowPriority"
- min_instances=0
- min_instances=1

```
        )
        cpu_cluster =
ml_client.begin_create_or_update(cpu_cluster)
)
```

Correct Answer:

Answer Area

```
from azure.ai.ml.entities import AmlCompute
try:
    cpu_cluster = ml_client.compute.get("cpu-cluster")
except Exception:
    cpu_cluster = [AmlCompute | ComputeInstance | KubernetesCompute] (
        name="cpu-cluster",
        size="STANDARD_DS3_V2",
        max_instances=4,
        [tier="LowPriority" | min_instances=0 | min_instances=1] ,
    )
    cpu_cluster =
ml_client.begin_create_or_update(cpu_cluster)
)
```

Question #101*Topic 2*

You create a workspace to include a compute instance by using Azure Machine Learning Studio. You are developing a Python SDK v2 notebook in the workspace.

You need to use Intellisense in the notebook.

What should you do?

- A. Stop the compute instance.
- B. Start the compute instance.

- C. Run a %pip magic function on the compute instance.
- D. Run a !pip magic function on the compute instance.

Correct Answer: *B*

Question #102*Topic 2*

DRAG DROP

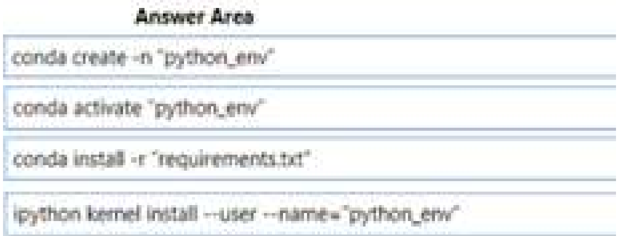
-

You manage an Azure Machine Learning workspace named workspace1 with a compute instance named compute1. You connect to compute1 by using a terminal window from workspace1. You create a file named "requirements.txt" containing Python dependencies to include Jupyter.

You need to add a new Jupyter kernel to compute1.

Which four commands should you use? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.

Correct Answer:



Question #103*Topic 2*

HOTSPOT

-

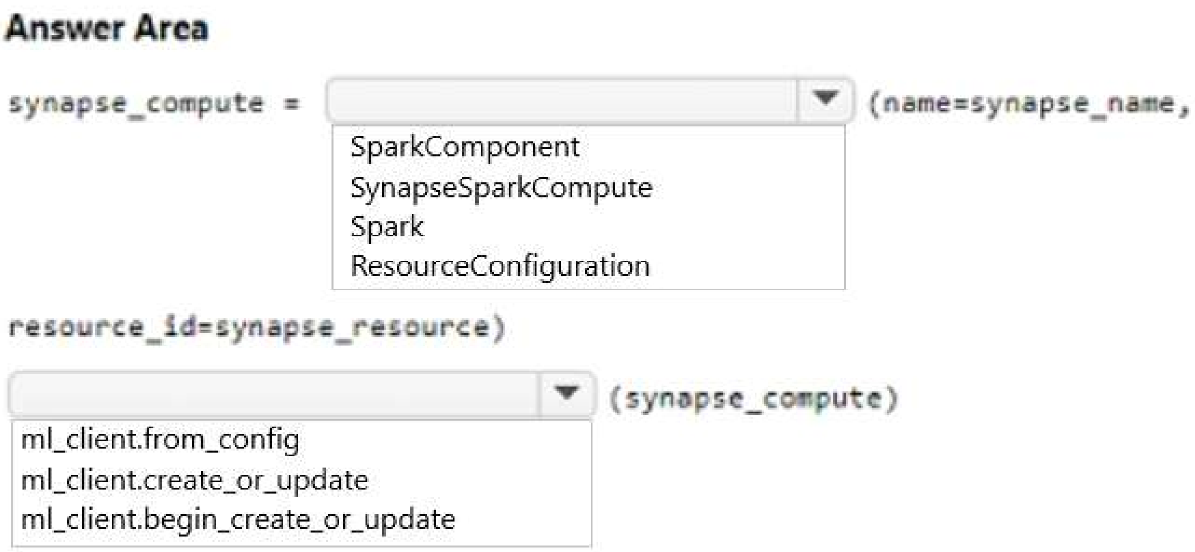
You are creating data wrangling and model training solutions in an Azure Machine Learning workspace.

You must use the same Python notebook to perform both data wrangling and model training.

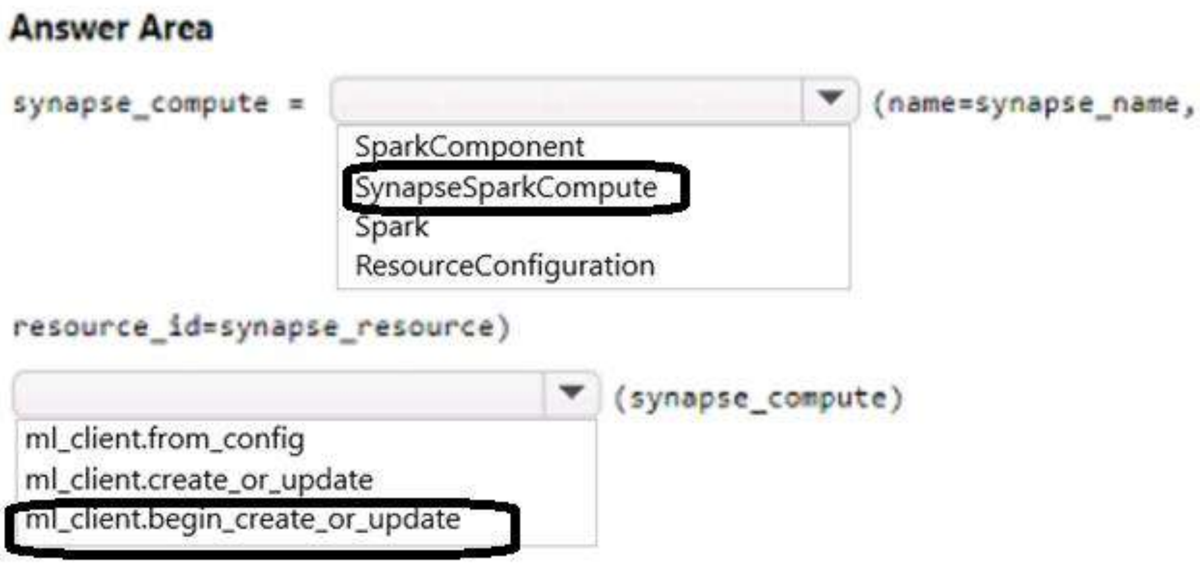
You need to use the Azure Machine Learning Python SDK v2 to define and configure the Synapse Spark pool asynchronously in the workspace as dedicated compute.

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.



Correct Answer:



Question #104*Topic 2*

HOTSPOT

-

You manage an Azure Machine Learning workspace named workspace1 by using the Python SDK v2.

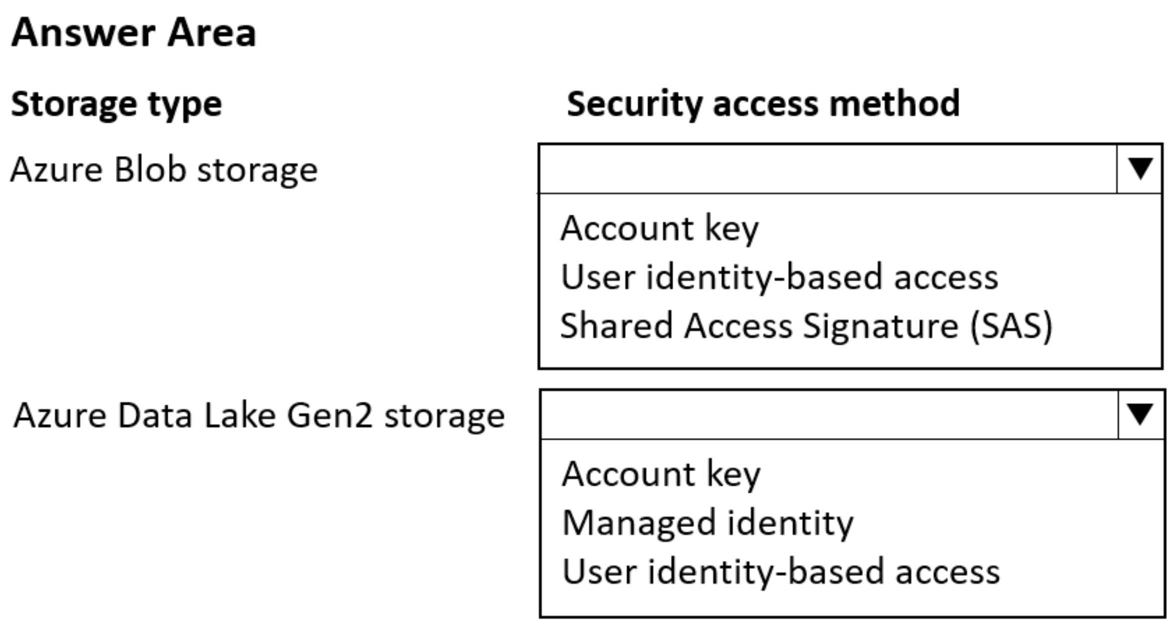
You must register datastores in workspace1 for Azure Blob and Azure Data Lake Gen2 storage to meet the following requirements:

• Data scientists accessing the datastore must have the same level of access.
• Access must be restricted to specified containers or folders.

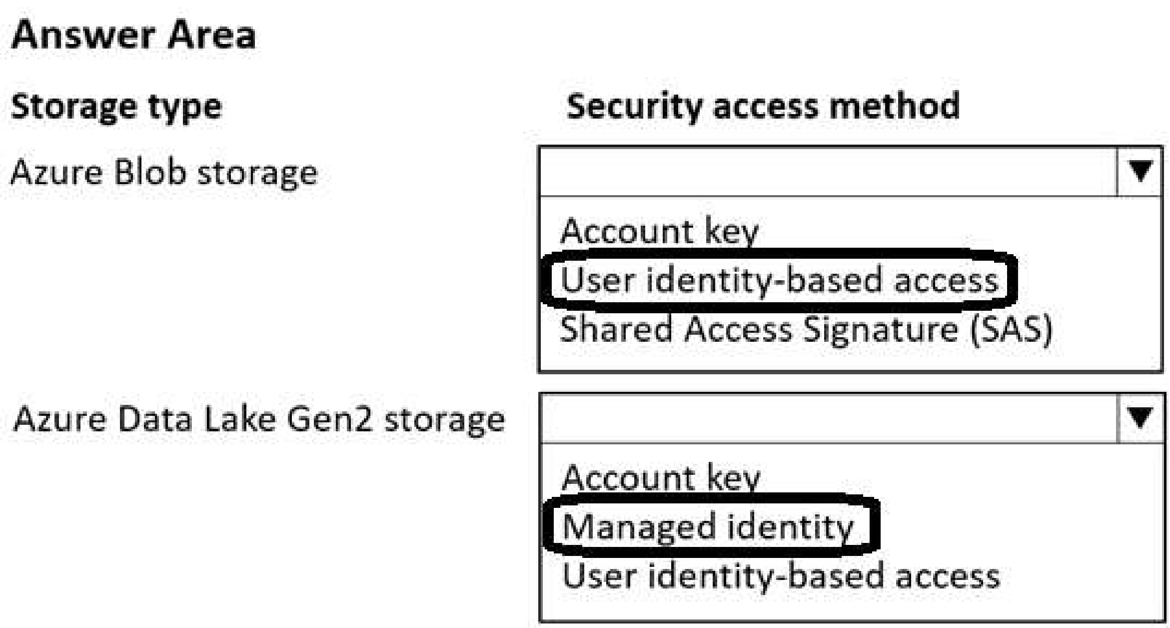
You need to configure a security access method used to register the Azure Blob and Azure Data Lake Gen2 storage in workspace1.

Which security access method should you configure? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.



Correct Answer:



Question #105*Topic 2*

DRAG DROP

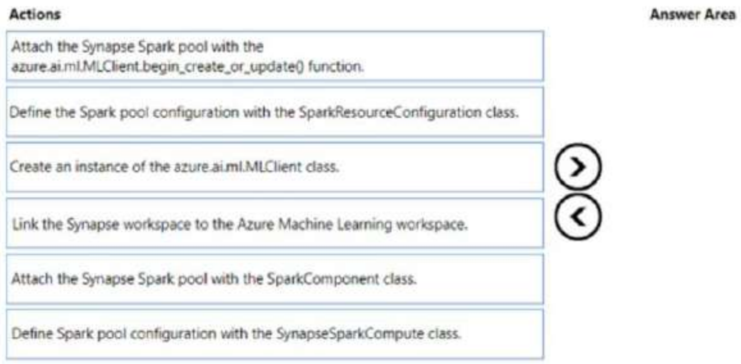
-

You create an Azure Machine Learning workspace and an Azure Synapse Analytics workspace with a Spark pool. The workspaces are contained within the same Azure subscription.

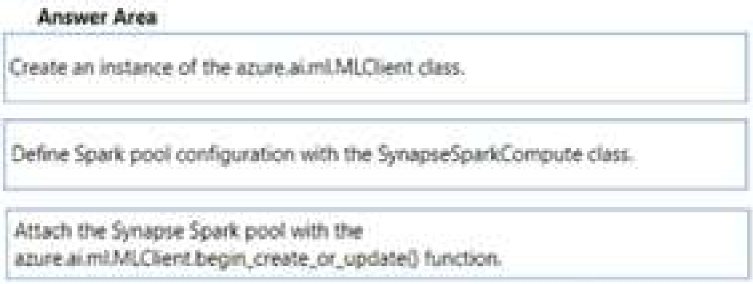
You must manage the Synapse Spark pool from the Azure Machine Learning workspace.

You need to attach the Synapse Spark pool in Azure Machine Learning by using the Python SDK v2.

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.



Correct Answer:



Question #106*Topic 2*

You manage an Azure Machine Learning workspace.

You need to define an environment from a Docker image by using the Azure Machine Learning Python SDK v2.

Which parameter should you use?

- A. properties
- B. image
- C. build
- D. conda_file

Correct Answer: *B*

Question #107*Topic 2*

You create an Azure Machine Learning managed compute resource. The compute resource is configured as follows:

• Minimum nodes: 2
• Maximum nodes: 4

You must decrease the minimum number of nodes and increase the maximum number of nodes to the following values:

• Minimum nodes: 0
• Maximum nodes: 8

You need to reconfigure the compute resource.

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

NOTE: Each correct selection is worth one point.

- A. Azure Machine Learning designer
- B. MLClient class in Python SDK v2
- C. Azure Machine Learning studio
- D. Azure CLI ml extension v2
- E. BuildContext class in Python SDK v2

Correct Answer: *ABD*

Question #108*Topic 2*

You plan to use automated machine learning by using Azure Machine Learning Python SDK v2 to train a regression model. You have data that has features with missing values, and categorical features with few distinct values.

You need to control whether automated machine learning automatically imputes missing values and encode categorical features as part of the training task.

Which enum of the automl package should you use?

- A. ForecastHorizonMode
- B. RegressionModels
- C. FeaturizationMode
- D. RegressionPrimaryMetrics

Correct Answer: *C*

Question #109*Topic 2*

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 use Azure Machine Learning designer to load the following datasets into an experiment:

Dataset1 -

Age	Length	Width
3	22	13
7	11	96
18	32	85

Dataset2 -

Age	Length	Width
11	101	65
6	98	23
33	22	54
17	52	12

You need to create a dataset that has the same columns and header row as the input datasets and contains all rows from both input datasets.

Solution: Use the Join Data module.

Does the solution meet the goal?

- A. Yes
- B. No

Correct Answer: *B*

Question #110*Topic 2*

HOTSPOT

-

You collect data from a nearby weather station. You have a pandas dataframe named weather_df that includes the following data:

Temperature	Observation_time	Humidity	Pressure	Visibility	Days_since_last observation
74	2019/10/2 00:00	0.62	29.87	3	0.5
89	2019/10/2 12:00	0.70	28.88	10	0.5
72	2019/10/3 00:00	0.64	30.00	8	0.5
80	2019/10/3 12:00	0.66	29.75	7	0.5

The data is collected every 12 hours: noon and midnight.

You plan to use automated machine learning to create a time-series model that predicts temperature over the next seven days. For the initial round of training, you want to train a maximum of 50 different models.

You must use the Azure Machine Learning SDK v2 to run an automated machine learning experiment to train these models.

You need to configure the automated machine learning job and its settings.

How should you configure parameters of the classes that implement the job? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Answer Area

job type
- regression
- forecasting
- classification
- deep learning

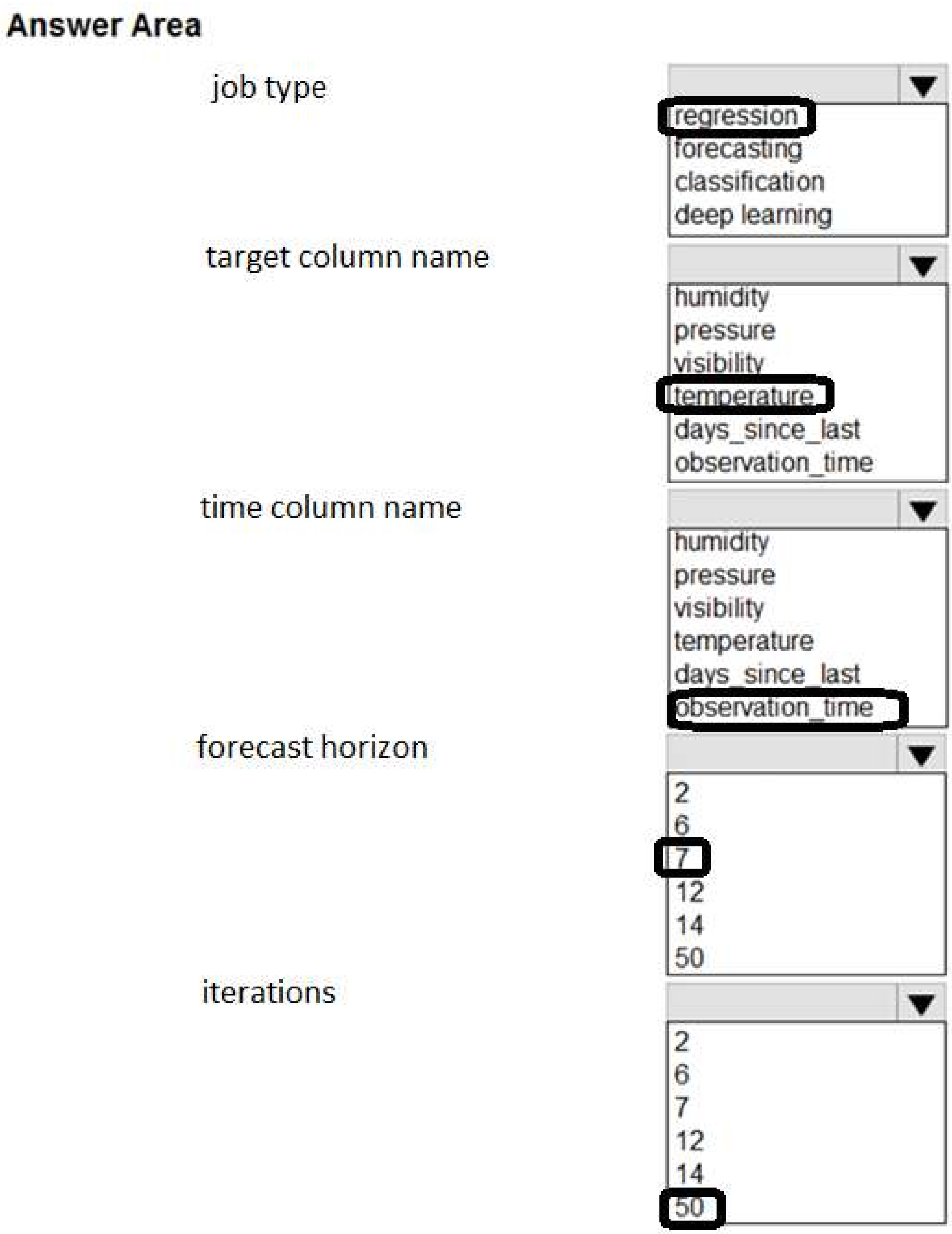
target column name
- humidity
- pressure
- visibility
- temperature
- days_since_last
- observation_time

time column name
- humidity
- pressure
- visibility
- temperature
- days_since_last
- observation_time

forecast horizon
- 2
- 6
- 7
- 12
- 14
- 50

iterations
- 2
- 6
- 7
- 12
- 14
- 50

Correct Answer:



Question #111*Topic 2*

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 have the following Azure subscriptions and Azure Machine Learning service workspaces:

Subscription	Workspace	Comment
385bdfe5-4cef-4ad4-b977-3f86d92727c9	ml-default	This is default subscription.
5a5891d1-557a-4234-9b83-2e90412b1068	ml-project	The information required to uniquely identify this workspace is stored in the file config.json in the same folder as the Python script.

You need to obtain a reference to the ml-project workspace.

Solution: Run the following Python code:

```
from azure.ai ml import MLClient
ws = MLClient.workspaces.get("ml-project")
```

Does the solution meet the goal?

- A. Yes
- B. No

Correct Answer: *B*

Question #112*Topic 2*
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 have the following Azure subscriptions and Azure Machine Learning service workspaces:

Subscription	Workspace	Comment
385bdfe5-4cef-4ad4-b977-3f86d92727c9	ml-default	This is default subscription.
5a5891d1-557a-4234-9b83-2e90412b1068	ml-project	The information required to uniquely identify this workspace is stored in the file config.json in the same folder as the Python script.

You need to obtain a reference to the ml-project workspace.

Solution: Run the following Python code:

```
from azure.ai.ml import MLClient
ws = MLClient(workspace_name= "ml-project")
```

Does the solution meet the goal?

- A. Yes
- B. No

Correct Answer: *B*

Question #113*Topic 2*

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 have the following Azure subscriptions and Azure Machine Learning service workspaces:

Subscription	Workspace	Comment
385bdfe5-4cef-4ad4-b977-3f86d92727c9	ml-default	This is default subscription.
5a5891d1-557a-4234-9b83-2e90412b1068	ml-project	The information required to uniquely identify this workspace is stored in the file config.json in the same folder as the Python script.

You need to obtain a reference to the ml-project workspace.

Solution: Run the following Python code:

```
from azure.ai.ml import MLClient
ws_from_config = MLClient.from_config()
```

Does the solution meet the goal?

- A. Yes
- B. No

Correct Answer: *B*

Question #114*Topic 2*

HOTSPOT
-

You manage an Azure Machine Learning workspace named workspace1.

You must register an Azure Blob storage datastore in workspace1 by using an access key. You develop Python SDK v2 code to import all modules required to register the datastore.

You need to complete the Python SDK v2 code to define the datastore.

How should you complete the code? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Answer Area

```
store = AzureBlobDatastore(
        name="blobdatastore",
        description="Sample blob datastore",
        account_name="sampleaccount1",
        [ ▼ ] ="samplecontainer1"
          filesystem
          container_name
          file_share_name
        protocol=" [ ▼ ] ",
                   abfss
                   wasbs
                   azureml
        credentials=AccountKeyCredentials{
                "account_key":
            "6L4VGVXLy30FM/RwmfLGppGj+xNetfUwDVBNPYwYDQDT1E+SirZcjP1a+xCHI
        +D9BTwLnYKhrERO+AStap8yKw=="
            },
    )
```

Correct Answer:

Answer Area

```
store = AzureBlobDatastore(
        name="blobdatastore",
        description="Sample blob datastore",
        account_name="sampleaccount1",
        [ ▼ ] ="samplecontainer1"
          filesystem
          container_name
          file_share_name
        protocol=" [ ▼ ] ",
                   abfss
                   wasbs
                   azureml
        credentials=AccountKeyCredentials{
                "account_key":
            "6L4VGVXLy30FM/RwmfLGppGj+xNetfUwDVBNPYwYDQDT1E+SirZcjP1a+xCHI
        +D9BTwLnYKhrERO+AStap8yKw=="
            },
    )
```

Question #115*Topic 2*

You manage an Azure Machine Learning workspace.

You plan to import data from Azure Data Lake Storage Gen2.

You need to build a URI that represents the storage location.

Which protocol should you use?

- A. https
- B. adl
- C. abfss
- D. wasbs

Correct Answer: *C*

Question #116*Topic 2*

You manage an Azure Machine Learning workspace. You have a folder that contains a CSV file. The folder is registered as a folder data asset.

You plan to use the folder data asset for data wrangling during interactive development.

You need to access and load the folder data asset into a Pandas data frame.

Which method should you use to achieve this goal?

- A. mltable.from_parquet_files()
- B. mltable.from_delimited_files()
- C. mltable.from_data_lake()
- D. mltable.load()

Correct Answer: *B*

Question #: 117
Topic #: 2

You manage an Azure Machine Learning workspace named proj1.

You plan to use assets defined in proj1 to create a pipeline in the Machine Learning studio designer.

You need to set the Registry name filter to display only the list of assets defined in proj1.

What should you set the Registry name filter to?

- A. proj1
- B. azureml-meta
- C. azureml
- D. workspace

Correct Answer: *A*

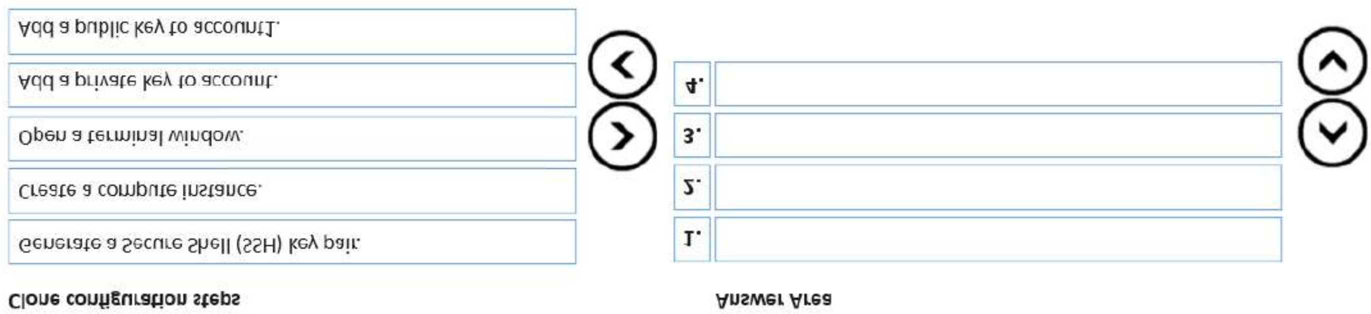
Question #: 118
Topic #: 2

DRAG DROP
-

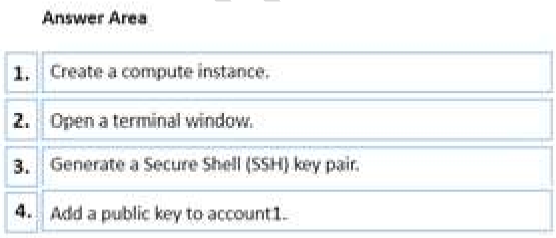
You have an Azure Machine Learning workspace named WS1 and a GitHub account named account1 that hosts a private repository named repo1.

You need to clone repo1 to make it available directly from WS1. The configuration must maximize the performance of the repo1 clone.

Which four actions should you perform in sequence?



Correct Answer:



Question #: 119
Topic #: 2

You manage an Azure Machine Learning workspace. You design a training job that is configured with a serverless compute.

The serverless compute must have a specific instance type and count.

You need to configure the serverless compute by using Azure Machine Learning Python SDK v2.

What should you do?

- A. Specify the compute name by using the compute parameter of the command job.
- B. Configure the tier parameter to Dedicated VM.
- C. Initialize and specify the ResourceConfiguration class.
- D. Initialize AmiCompute class with size and type specification.

Correct Answer: *C*

Question #: 120
Topic #: 2

HOTSPOT
-

You manage an Azure subscription that contains the following resources:

Name	Resource type	Description
Workspace1	Azure Machine Learning workspace	Contains a training pipeline named **MLPipeline1** and a dataset named **Dataset1**
DataFactory1	Azure Data Factory	Contains a pipeline named **DFPipeline1** that runs **MLPipeline1**

You plan to implement a solution that will automatically trigger the retraining of the model implemented by MLPipeline1. The trigger must be invoked if data drift is detected in Dataset1.

You need to select the components to invoke and run the solution. The solution must minimize coding implementation and maintenance efforts.

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

NOTE: Each correct selection is worth one point.

Answer Area

Retraining triggers

Component type	Value
Component used in invoke	Azure Event Hubs topic Azure Event Grid subscription Azure Machine Learning linked service
Component used to run	Azure Logic app Azure Functions app Azure Automation runbook

Correct Answer:

Answer Area

Retraining triggers

Component type	Value
Component used in invoke	Azure Event Hubs topic Azure Event Grid subscription Azure Machine Learning linked service
Component used to run	Azure Logic app Azure Functions app Azure Automation runbook

Question #: 121
Topic #: 2

You manage an Azure Machine Learning workspace.

You must create and configure a compute cluster for a training job by using Python SDK v2.

You need to create a persistent Azure Machine Learning compute resource, specifying the fewest possible properties.

Which two properties should you define? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

- A. size
- B. win_instances
- C. type
- D. name
- E. max_instances

Correct Answer: *AE*

Question #: 122
Topic #: 2

You manage an Azure Machine Learning workspace named Workspace1.

You plan to create a pipeline in the Azure Machine Learning Studio designer. The pipeline must include a custom component.

You need to ensure the custom component can be used in the pipeline.

What should you do first?

- A. Create a pipeline endpoint.
- B. Add a linked service to Workspace1.
- C. Upload a .json file to Workspace1.
- D. Create a datastore.
- E. Upload a .yaml file to Workspace1.

Correct Answer: *E*

Question #: 123
Topic #: 2

You have an Azure Machine Learning workspace and a serverless Spark compute resource.

You plan to run the same Spark session every 40 minutes. Each session runs for 15 minutes.

During testing, you observe a 15-minute delay at the start of the Spark sessions.

You need to reduce the delay to less than 1 minute.

What should you do?

- A. Configure the session timeout to be 25 minutes.
- B. Increase the number of nodes to 16.
- C. Enable dynamically allocated executors.
- D. Enable an isolated compute.

Correct Answer: *A*

Question #: 124
Topic #: 2

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 manage an Azure Machine Learning workspace. The development environment for managing the workspace is configured to use Python SDK v2 in Azure Machine Learning Notebooks.

A Synapse Spark Compute is currently attached and uses system-assigned identity.

You need to use Python code to update the Synapse Spark Compute to use a user-assigned identity.

Solution: Create an instance of the MLClient class.

Does the solution meet the goal?

- A. Yes
- B. No

Correct Answer: *B*

Question #: 125
Topic #: 2

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 manage an Azure Machine Learning workspace. The development environment for managing the workspace is configured to use Python SDK v2 in Azure Machine Learning Notebooks.

A Synapse Spark Compute is currently attached and uses system-assigned identity.

You need to use Python code to update the Synapse Spark Compute to use a user-assigned identity.

Solution: Initialize the DefaultAzureCredential class.

Does the solution meet the goal?

- A. Yes
- B. No

Correct Answer: *B*

Question #: 126
Topic #: 2

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 manage an Azure Machine Learning workspace. The development environment for managing the workspace is configured to use Python SDK v2 in Azure Machine Learning Notebooks.

A Synapse Spark Compute is currently attached and uses system-assigned identity.

You need to use Python code to update the Synapse Spark Compute to use a user-assigned identity.

Solution: Configure the IdentityConfiguration class with the appropriate identity type.

Does the solution meet the goal?

- A. Yes
- B. No

Correct Answer: *A*

Question #: 127
Topic #: 2

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 have an Azure Machine Learning workspace that includes an AmlCompute cluster and a batch endpoint.

You clone a repository that contains an MLflow model to your local computer.

You need to ensure that you can deploy the model to the batch endpoint.

Solution: Add a compute resource to the workspace.

Does the solution meet the goal?

- A. Yes
- B. No

Correct Answer: *A*

Question #: 128
Topic #: 2

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 have an Azure Machine Learning workspace that includes an AmlCompute cluster and a batch endpoint.

You clone a repository that contains an MLflow model to your local computer.

You need to ensure that you can deploy the model to the batch endpoint.

Solution: Register the model in the workspace.

Does the solution meet the goal?

- A. Yes
- B. No

Correct Answer: *B*

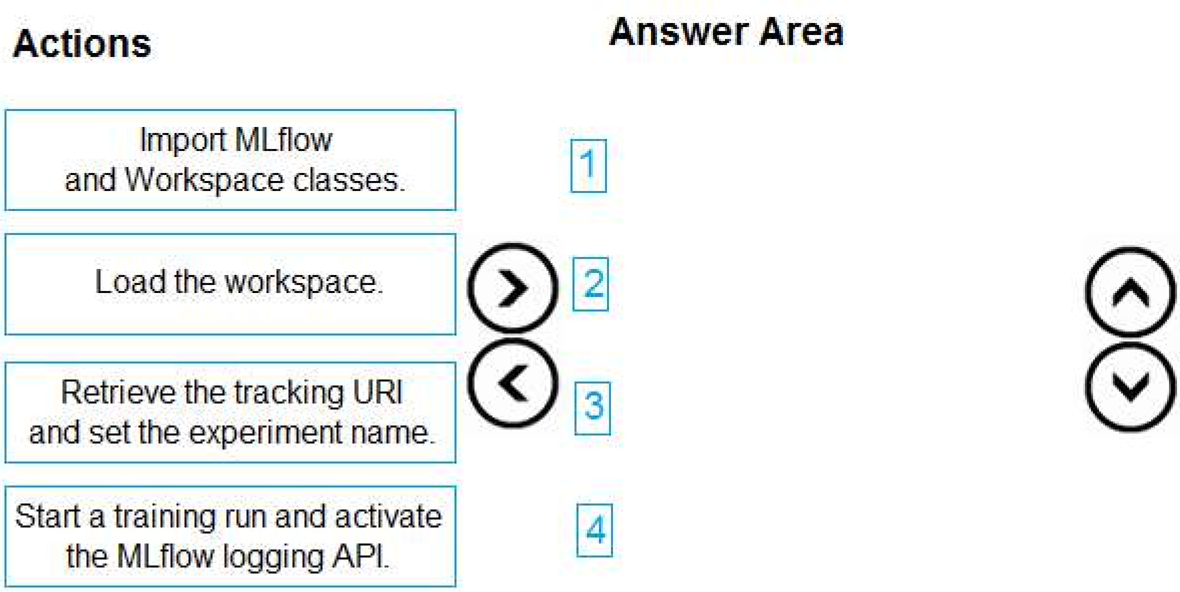
Question #: 129
Topic #: 3

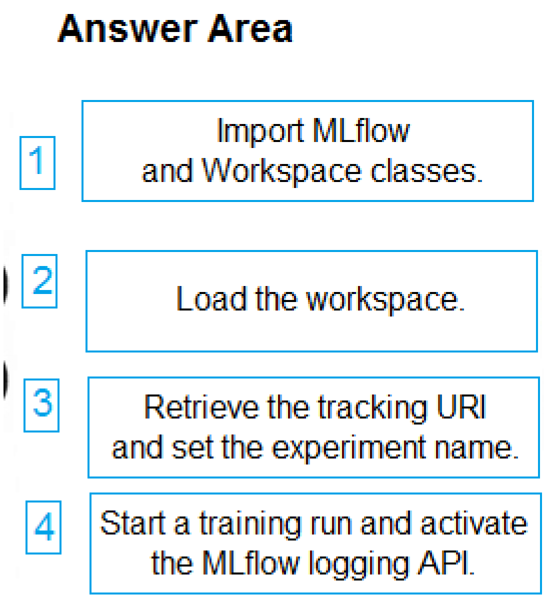
DRAG DROP
-

You have an Azure Machine Learning workspace. You are running an experiment on your local computer.

You need to use MLflow Tracking to store metrics and artifacts from your local experiment runs in the workspace.

In which order should you perform the actions? To answer, move all actions from the list of actions to the answer area and arrange them in the correct order.





Correct Answer:

Question #: 130
Topic #: 2

HOTSPOT
-

You design a data processing strategy for a machine learning project.

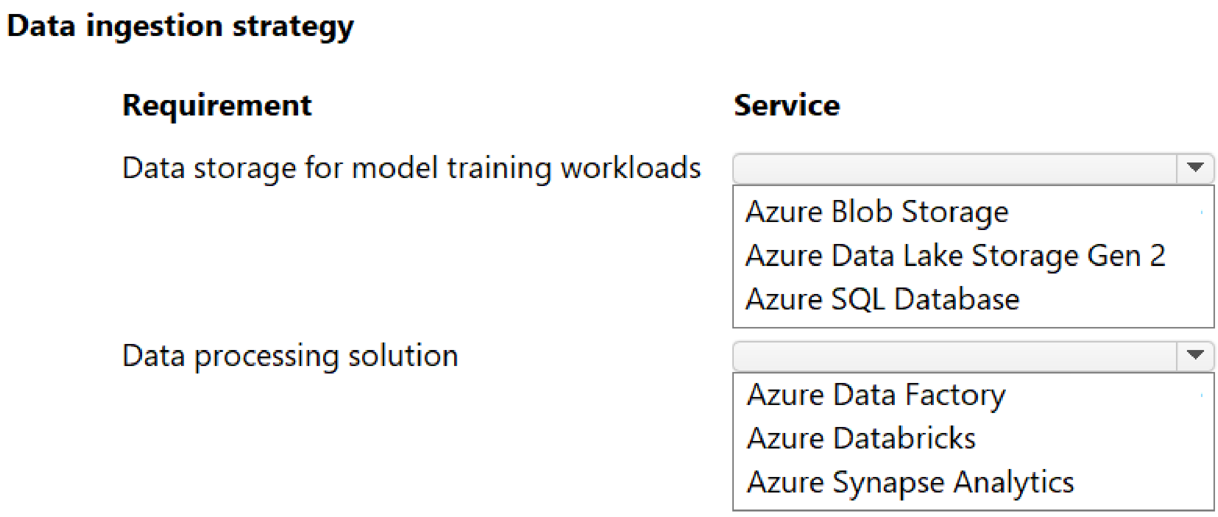
The data that must be processed includes unstructured flat files that must be processed in real time.

The data transformation must be executed on a serverless compute and optimized for big data analytical workloads.

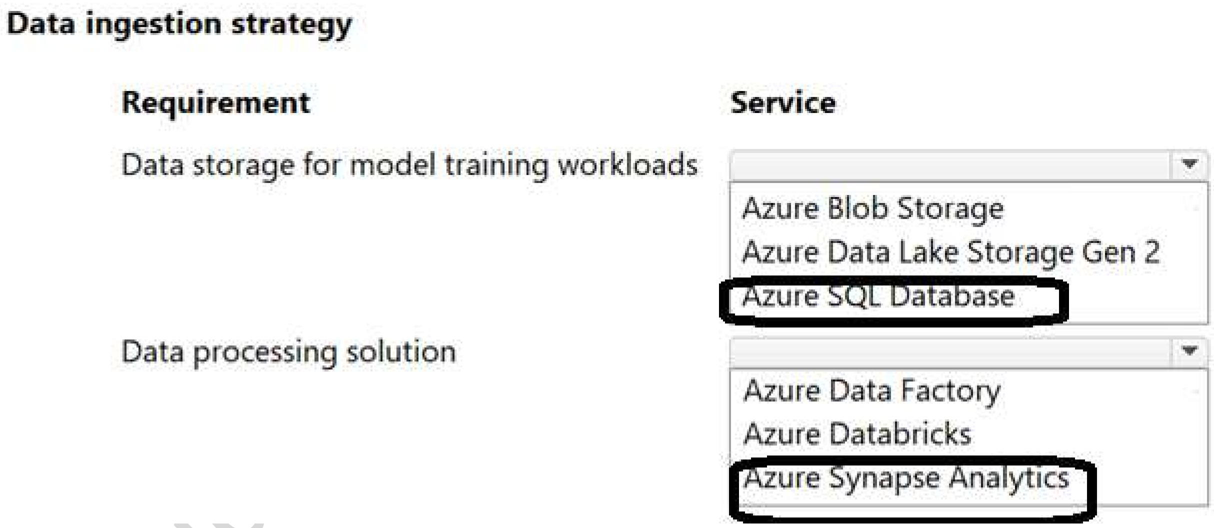
You need to select the Azure services for the data science team.

Which storage and data processing service should you use? To answer, select the appropriate option in the answer area.

NOTE: Each correct selection is worth one point.



Correct Answer:



Question #: 131
Topic #: 2

DRAG DROP
-

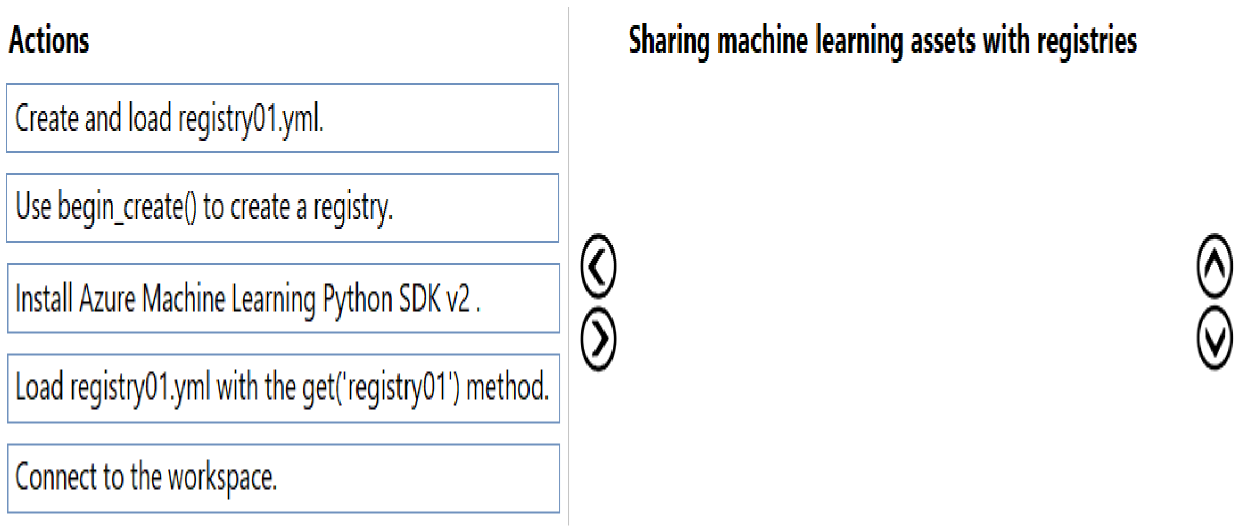
You manage an Azure Machine Learning workspace named workspace1.

You plan to create a registry named registry01 with the help of the following registry.yml (line numbers are used for reference only):

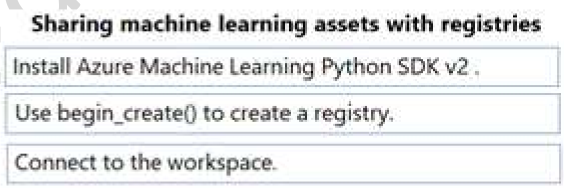
```
name: DemoRegistry1
tags:
  description: A registry with one primary region and
  foo: bar
location: eastus
replication_locations:
  - location: eastus
  - location: eastus2
```

You need to use Azure Machine Learning Python SDK v2 with Python 3.10 in a notebook to interact with workspace1.

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.



Correct Answer:



Question #: 132
Topic #: 2

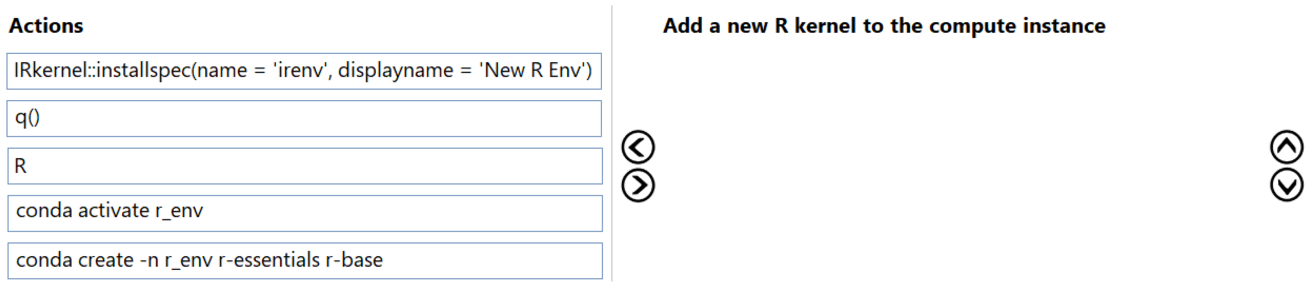
DRAG DROP
-

You have an Azure Machine Learning workspace.

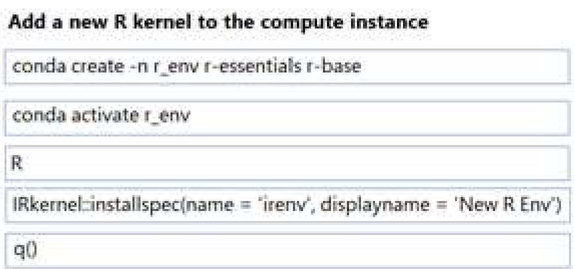
You plan to use the terminal to configure a compute instance to run a notebook.

You need to add a new R kernel to the compute instance.

In which order should you perform the actions? To answer, move all actions from the list of actions to the answer area and arrange them in the correct order.



Correct Answer:



Question #: 133
Topic #: 2

HOTSPOT
-

You manage an Azure Machine Learning workspace named Workspace1 and an Azure Blob Storage accessed by using the URL https://storage1.blob.core.windows.net/data1.

You plan to create an Azure Blob datastore in Workspace1. The datastore must target the Blob Storage by using Azure Machine Learning Python SDK v2. Access authorization to the datastore must be limited to a specific amount of time.

You need to select the parameters of the AzureBlobDatastore class that will point to the target datastore and authorize access to it.

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

NOTE: Each correct selection is worth one point.

Datastore parameters

Parameter purpose	Value
Point to the target datastore.	container_name="data1" filesystem="data1" file_share_name="data1"
Authorize access.	credentials=AccountKeyConfiguration credentials=SaSTokenConfiguration credentials=ServicePrincipalCredentials

Correct Answer:

Datastore parameters

Parameter purpose	Value
Point to the target datastore.	container_name="data1" filesystem="data1" file_share_name="data1"
Authorize access.	credentials=AccountKeyConfiguration credentials=SaSTokenConfiguration credentials=ServicePrincipalCredentials

Question #: 134
Topic #: 2

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 manage an Azure Machine Learning workspace. The development environment for managing the workspace is configured to use Python SDK v2 in Azure Machine Learning Notebooks.

A Synapse Spark Compute is currently attached and uses system-assigned identity.

You need to use Python code to update the Synapse Spark Compute to use a user-assigned identity.

Solution: Pass the UserAssignedIdentity class object to the SynapseSparkCompute class.

Does the solution meet the goal?

- A. Yes
- B. No

Correct Answer: *A*

Question #: 135
Topic #: 2

DRAG DROP
-

You have an Azure Machine Learning workspace named WS1.

You plan to use WS1 to train two models named model1 and model2. For model1, you plan to use automated machine learning. For model2, you plan to use Azure Machine Learning designer.

You need to determine the compute targets you should use to train each model. Your solution must ensure the following:

• The compute target for model1 supports auto-shutdown/auto-start based on a schedule.
• The compute target for model2 supports the use of low-priority Azure Virtual Machines.

Which compute targets should you use? To answer, move the appropriate compute targets to the correct model. You may use each compute target once, more than once, or not at all. You may need to move the split bar between panes or scroll to view content.

NOTE: Each correct selection is worth one point.

Compute targets

- Azure Machine Learning Kubernetes
- Azure Machine Learning compute cluster
- Azure Machine Learning compute instance

Models and compute targets

Model	Compute target
model1	
model2	

Correct Answer:

Models and compute targets

Model	Compute target
model1	Azure Machine Learning compute instance
model2	Azure Machine Learning compute cluster

Topic 3 - Question Set 3

Question #1*Topic 3*

You are analyzing a dataset containing historical data from a local taxi company. You are developing 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. a Root Mean Square Error value that is low
- B. an R-Squared value close to 0
- C. an F1 score that is low
- D. an R-Squared value close to 1
- E. an F1 score that is high
- F. a Root Mean Square Error value that is high

Correct Answer: *AD*
RMSE and R2 are both metrics for regression models.
A: 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.
D: 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.
Incorrect Answers:
C, E: F-score is used for classification models, not for regression models.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/evaluate-model

Question #2*Topic 3*

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 to run an experiment that trains a classification model.
You want to use Hyperdrive to find parameters that optimize the AUC metric for the model. You configure a HyperDriveConfig for the experiment by running the following code:

```
hyperdrive = HyperDriveConfig(estimator=your_estimator,
 hyperparameter_sampling=your_params,
 policy=policy,
 primary_metric_name='AUC',
 primary_metric_goal=PrimaryMetricGoal.MAXIMIZE,
 max_total_runs=6,
 max_concurrent_runs=4)
```

You plan to use this configuration to run a script that trains a random forest model and then tests it with validation data. The label values for the validation data are stored in a variable named y_test variable, and the predicted probabilities from the model are stored in a variable named y_predicted.
You need to add logging to the script to allow Hyperdrive to optimize hyperparameters for the AUC metric.
Solution: Run the following code:

```
from sklearn.metrics import roc_auc_score
import logging
# code to train model omitted
auc = roc_auc_score(y_test, y_predicted)
logging.info("AUC: " + str(auc))
```

Does the solution meet the goal?

- A. Yes
- B. No

Correct Answer: *B*
Python printing/logging example:
logging.info(message)
Destination: Driver logs, Azure Machine Learning designer
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-debug-pipelines

Question #3*Topic 3*

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 to run an experiment that trains a classification model.
You want to use Hyperdrive to find parameters that optimize the AUC metric for the model. You configure a HyperDriveConfig for the experiment by running the following code:

```
hyperdrive = HyperDriveConfig(estimator=your_estimator,
 hyperparameter_sampling=your_params,
 policy=policy,
 primary_metric_name='AUC',
 primary_metric_goal=PrimaryMetricGoal.MAXIMIZE,
 max_total_runs=6,
 max_concurrent_runs=4)
```

You plan to use this configuration to run a script that trains a random forest model and then tests it with validation data. The label values for the validation data are stored in a variable named y_test variable, and the predicted probabilities from the model are stored in a variable named y_predicted. You need to add logging to the script to allow Hyperdrive to optimize hyperparameters for the AUC metric.

Solution: Run the following code:

```
import json, os
from sklearn.metrics import roc_auc_score
# code to train model omitted
auc = roc_auc_score(y_test, y_predicted)
os.makedirs("outputs", exist_ok = True)
with open("outputs/AUC.txt", "w") as file_cur:
    file_cur.write(auc)
```

Does the solution meet the goal?

- A. Yes
- B. No

Correct Answer: *B*

Explanation -

Use a solution with logging.info(message) instead.

Note: Python printing/logging example:

logging.info(message)

Destination: Driver logs, Azure Machine Learning designer

Reference:

https://docs.microsoft.com/en-us/azure/machine-learning/how-to-debug-pipelines

Question #4*Topic 3*

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 to run an experiment that trains a classification model.

You want to use Hyperdrive to find parameters that optimize the AUC metric for the model. You configure a HyperDriveConfig for the experiment by running the following code:

```
hyperdrive = HyperDriveConfig(estimator=your_estimator,
 hyperparameter_sampling=your_params,
 policy=policy,
 primary_metric_name='AUC',
 primary_metric_goal=PrimaryMetricGoal.MAXIMIZE,
 max_total_runs=6,
 max_concurrent_runs=4)
```

You plan to use this configuration to run a script that trains a random forest model and then tests it with validation data. The label values for the validation data are stored in a variable named y_test variable, and the predicted probabilities from the model are stored in a variable named y_predicted.
You need to add logging to the script to allow Hyperdrive to optimize hyperparameters for the AUC metric.
Solution: Run the following code:

```
import numpy as np
from sklearn.metrics import roc_auc_score
# code to train model omitted
auc = roc_auc_score(y_test, y_predicted)
print(np.float(auc))
```

Does the solution meet the goal?

- A. Yes
- B. No

Correct Answer: *B*
Explanation -
Use a solution with logging.info(message) instead.
Note: Python printing/logging example:
logging.info(message)
Destination: Driver logs, Azure Machine Learning designer
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-debug-pipelines

Question #5*Topic 3*

You use the following code to run a script as an experiment in Azure Machine Learning:

```
from azureml.core import Workspace, Experiment, Run
from azureml.core import RunConfig, ScriptRunConfig
ws = Workspace.from_config()
run_config = RunConfiguration()
run_config.target='local'
script_config = ScriptRunConfig(source_directory='./script', script='experiment.py', run_config=run_config)
experiment = Experiment(workspace=ws, name='script experiment')
run = experiment.submit(config=script_config)
run.wait_for_completion()
```

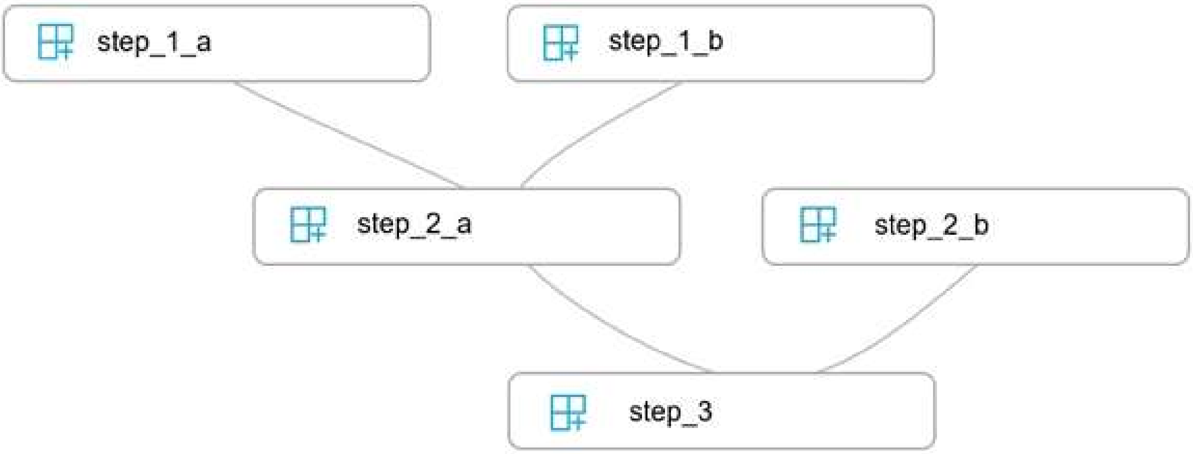
You must identify the output files that are generated by the experiment run.
You need to add code to retrieve the output file names.
Which code segment should you add to the script?

- A. files = run.get_properties()
- B. files= run.get_file_names()
- C. files = run.get_details_with_logs()
- D. files = run.get_metrics()
- E. files = run.get_details()

Correct Answer: *B*
You can list all of the files that are associated with this run record by called run.get_file_names()
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-track-experiments

Question #6*Topic 3*

You write five Python scripts that must be processed in the order specified in Exhibit A `" which allows the same modules to run in parallel, but will wait for modules with dependencies.
You must create an Azure Machine Learning pipeline using the Python SDK, because you want to script to create the pipeline to be tracked in your version control system. You have created five PythonScriptSteps and have named the variables to match the module names.



You need to create the pipeline shown. Assume all relevant imports have been done.
Which Python code segment should you use?
A.

```
p = Pipeline(ws, steps=[[[[step_1_a, step_1_b], step_2_a], step_2_b], step_3])
```

B.

```
pipeline_steps = {
    "Pipeline": {
        "run": step_3,
        "run_after": {[
            {"run": step_2_a,
                "run_after":
                    [{"run": step_1_a},
                     {"run": step_1_b}]
                },
            {"run": step_2_b}]
        }
    }
}
p = Pipeline(ws, steps=pipeline_steps)
```

C.

```
step_2_a.run_after(step_1_b)
step_2_a.run_after(step_1_a)
step_3.run_after(step_2_b)
step_3.run_after(step_2_a)
p = Pipeline(ws, steps=[step_3])
```

D.

```
p = Pipeline(ws, steps=[step_1_a, step_1_b, step_2_a, step_2_b, step_3])
```

Correct Answer: *A*
The steps parameter is an array of steps. To build pipelines that have multiple steps, place the steps in order in this array.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-use-parallel-run-step
Question #7*Topic 3*

You create a datastore named training_data that references a blob container in an Azure Storage account. The blob container contains a folder named csv_files in which multiple comma-separated values (CSV) files are stored.
You have a script named train.py in a local folder named ./script that you plan to run as an experiment using an estimator. The script includes the following code to read data from the csv_files folder:

```
import os
import argparse
import pandas as pd

from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from azureml.core import Run

run = Run.get_context()
parser = argparse.ArgumentParser()
parser.add_argument('--data-folder', type=str, dest='data_folder', help='data reference')
args = parser.parse_args()

data_folder = args.data_folder
csv_files = os.listdir(data_folder)
training_data = pd.concat((pd.read_csv(os.path.join(data_folder,csv_file)) for csv_file in csv_files))

# Code goes on to split the training data and train a logistic regression model
```

You have the following script.

```
from azureml.core import Workspace, Datastore, Experiment
from azureml.train.sklearn import SKLearn

ws = Workspace.from_config()
exp = Experiment(workspace=ws, name='csv_training')
ds = Datastore.get(ws, datastore_name='training_data')
data_ref = ds.path('csv_files')

# Code to define estimator goes here

run = exp.submit(config=estimator)
run.wait_for_completion(show_output=True)
```

You need to configure the estimator for the experiment so that the script can read the data from a data reference named data_ref that references the csv_files folder in the training_data datastore. Which code should you use to configure the estimator?

A.

```
estimator = SKLearn(source_directory='./script',
 inputs=[data_ref.as_named_input('data-folder').to_pandas_dataframe()],
 compute_target='local',
 entry_script='train.py')
```

B.

```
script_params = {
   '--data-folder': data_ref.as_mount()
}
estimator = SKLearn(source_directory='./script',
 script_params=script_params,
 compute_target='local',
 entry_script='train.py'
```

C.

```
estimator = SKLearn(source_directory='./script',
 inputs=[data_ref.as_named_input('data-folder').as_mount()],
 compute_target='local',
 entry_script='train.py')
```

D.

```
script_params = {
    '--data-folder': data_ref.as_download(path_on_compute='csv_files')
}
estimator = SKLearn(source_directory='./script',
 script_params=script_params,
 compute_target='local',
 entry_script='train.py'
```

E.

```
estimator = SKLearn(source_directory='./script',
 inputs=[data_ref.as_named_input('data-folder').as_download(path_on_compute='csv_files')],
 compute_target='local',
 entry_script='train.py')
```

Correct Answer: *B*
Besides passing the dataset through the input parameters in the estimator, you can also pass the dataset through script_params and get the data path (mounting point) in your training script via arguments. This way, you can keep your training script independent of azureml-sdk. In other words, you will be able use the same training script for local debugging and remote training on any cloud platform.
Example:
from azureml.train.sklearn import SKLearn
script_params = {
# mount the dataset on the remote compute and pass the mounted path as an argument to the training script
'--data-folder': mnist_ds.as_named_input('mnist').as_mount(),
'--regularization': 0.5
}
est = SKLearn(source_directory=script_folder,
script_params=script_params,
compute_target=compute_target,
environment_definition=env,
entry_script='train_mnist.py')
# Run the experiment
run = experiment.submit(est)
run.wait_for_completion(show_output=True)
Incorrect Answers:
A: Pandas DataFrame not used.
Reference:
https://docs.microsoft.com/es-es/azure/machine-learning/how-to-train-with-datasets
Question #8*Topic 3*

DRAG DROP -
You create a multi-class image classification deep learning experiment by using the PyTorch framework. You plan to run the experiment on an Azure Compute cluster that has nodes with GPU's.
You need to define an Azure Machine Learning service pipeline to perform the monthly retraining of the image classification model. The pipeline must run with minimal cost and minimize the time required to train the model.
Which three pipeline steps 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.

Select and Place:

Actions

- Configure a DataTransferStep() to fetch new image data from public web portal, running on the cpu-compute compute target.
- Configure an EstimatorStep() to run an estimator that runs the bird_classifier_train.py model training script on the gpu_compute compute target.
- Configure a PythonScriptStep() to run both image_fetcher.py and image_resize.py on the cpu-compute compute target.
- Configure an EstimatorStep() to run an estimator that runs the bird_classifier_train.py model training script on the cpu_compute compute target.
- Configure a PythonScriptStep() to run image_fetcher.py on the cpu-compute compute target.
- Configure a PythonScriptStep() to run image_resize.py on the cpu-compute compute target.
- Configure a PythonScriptStep() to run bird_classifier_train.py on the cpu-compute compute target.
- Configure a PythonScriptStep() to run bird_classifier_train.py on the gpu-compute compute target.

Answer Area

Correct Answer:

Actions	Answer Area
Configure a DataTransferStep() to fetch new image data from public web portal, running on the cpu-compute compute target.	Configure a DataTransferStep() to fetch new image data from public web portal, running on the cpu-compute compute target.
Configure an EstimatorStep() to run an estimator that runs the bird_classifier_train.py model training script on the gpu_compute compute target.	Configure a PythonScriptStep() to run image_resize.py on the cpu-compute compute target.
Configure a PythonScriptStep() to run both image_fetcher.py and image_resize.py on the cpu-compute compute target.	Configure an EstimatorStep() to run an estimator that runs the bird_classifier_train.py model training script on the gpu_compute compute target.
Configure an EstimatorStep() to run an estimator that runs the bird_classifier_train.py model training script on the cpu_compute compute target.	
Configure a PythonScriptStep() to run image_fetcher.py on the cpu-compute compute target.	
Configure a PythonScriptStep() to run image_resize.py on the cpu-compute compute target.	
Configure a PythonScriptStep() to run bird_classifier_train.py on the cpu-compute compute target.	
Configure a PythonScriptStep() to run bird_classifier_train.py on the gpu-compute compute target.	

Step 1: Configure a DataTransferStep() to fetch new image dataâ€¦
Step 2: Configure a PythonScriptStep() to run image_resize.y on the cpu-compute compute target.
Step 3: Configure the EstimatorStep() to run training script on the gpu_compute computer target.
The PyTorch estimator provides a simple way of launching a PyTorch training job on a compute target.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-train-pytorch
Question #9*Topic 3*

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.
An IT department creates the following Azure resource groups and resources:

Resource group	Resources
ml_resources	• an Azure Machine Learning workspace named amlworkspace • an Azure Storage account named amlworkspace12345 • an Application Insights instance named amlworkspace54321 • an Azure Key Vault named amlworkspace67890 • an Azure Container Registry named amlworkspace09876
general_compute	A virtual machine named mlvm with the following configuration: • Operating system: Ubuntu Linux • Software installed: Python 3.6 and Jupyter Notebooks • Size: NC6 (6 vCPUs, 1 vGPU, 56 Gb RAM)

The IT department creates an Azure Kubernetes Service (AKS)-based inference compute target named aks-cluster in the Azure Machine Learning workspace.

You have a Microsoft Surface Book computer with a GPU. Python 3.6 and Visual Studio Code are installed.
You need to run a script that trains a deep neural network (DNN) model and logs the loss and accuracy metrics.
Solution: Attach the mlvm virtual machine as a compute target in the Azure Machine Learning workspace. Install the Azure ML SDK on the Surface Book and run
Python code to connect to the workspace. Run the training script as an experiment on the mlvm remote compute resource.
Does the solution meet the goal?

- A. Yes
- B. No

Correct Answer: *A*
Use the VM as a compute target.
Note: A compute target is a designated compute resource/environment where you run your training script or host your service deployment. This location may be your local machine or a cloud-based compute resource.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/concept-compute-target

Question #10*Topic 3*

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.
An IT department creates the following Azure resource groups and resources:

Resource group	Resources
ml_resources	• an Azure Machine Learning workspace named amlworkspace • an Azure Storage account named amlworkspace12345 • an Application Insights instance named amlworkspace54321 • an Azure Key Vault named amlworkspace67890 • an Azure Container Registry named amlworkspace09876
general_compute	A virtual machine named mlvm with the following configuration: • Operating system: Ubuntu Linux • Software installed: Python 3.6 and Jupyter Notebooks • Size: NC6 (6 vCPUs, 1 vGPU, 56 Gb RAM)

The IT department creates an Azure Kubernetes Service (AKS)-based inference compute target named aks-cluster in the Azure Machine Learning workspace.
You have a Microsoft Surface Book computer with a GPU. Python 3.6 and Visual Studio Code are installed.
You need to run a script that trains a deep neural network (DNN) model and logs the loss and accuracy metrics.
Solution: Install the Azure ML SDK on the Surface Book. Run Python code to connect to the workspace and then run the training script as an experiment on local compute.
Does the solution meet the goal?

- A. Yes

- B. No

Correct Answer:A
Need to attach the mlvm virtual machine as a compute target in the Azure Machine Learning workspace.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/concept-compute-target

Question #11*Topic 3*

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.
An IT department creates the following Azure resource groups and resources:

Resource group	Resources
ml_resources	• an Azure Machine Learning workspace named amlworkspace • an Azure Storage account named amlworkspace12345 • an Application Insights instance named amlworkspace54321 • an Azure Key Vault named amlworkspace67890 • an Azure Container Registry named amlworkspace09876
general_compute	A virtual machine named mlvm with the following configuration: • Operating system: Ubuntu Linux • Software installed: Python 3.6 and Jupyter Notebooks • Size: NC6 (6 vCPUs, 1 vGPU, 56 Gb RAM)

The IT department creates an Azure Kubernetes Service (AKS)-based inference compute target named aks-cluster in the Azure Machine Learning workspace.
You have a Microsoft Surface Book computer with a GPU. Python 3.6 and Visual Studio Code are installed.
You need to run a script that trains a deep neural network (DNN) model and logs the loss and accuracy metrics.
Solution: Install the Azure ML SDK on the Surface Book. Run Python code to connect to the workspace. Run the training script as an experiment on the aks- cluster compute target.
Does the solution meet the goal?

- A. Yes
- B. No

Correct Answer: *B*
Need to attach the mlvm virtual machine as a compute target in the Azure Machine Learning workspace.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/concept-compute-target

Question #12*Topic 3*

HOTSPOT -
You plan to use Hyperdrive to optimize the hyperparameters selected when training a model. You create the following code to define options for the hyperparameter experiment:

```
import azureml.train.hyperdrive.parameter_expressions as pe
from azureml.train.hyperdrive import GridParameterSampling, HyperDriveConfig

param_sampling = GridParameterSampling({
  "max_depth" : pe.choice(6, 7, 8, 9),
  "learning_rate" : pe.choice(0.05, 0.1, 0.15)
  })
hyperdrive_run_config = HyperDriveConfig(
  estimator = estimator,
  hyperparameter_sampling = param_sampling,
  policy = None,
  primary_metric_name = "auc",
  primary_metruc_goal = PrimaryMetricGoal.MAXIMIZE,
  max_total_runs = 50,
  max_concurrent_runs = 4)
```

For each of the following statements, select Yes if the statement is true. Otherwise, select No.
NOTE: Each correct selection is worth one point.
Hot Area:

Answer Area

	Yes	No
There will be 50 runs for this hyperparameter tuning experiment.	○	○
You can use the policy parameter in the HyperDriveConfig class to specify a security policy.	○	○
The experiment will create a run for every possible value for the learning rate parameter between 0.05 and 0.15.	○	○

Correct Answer:

Answer Area

	Yes	No
There will be 50 runs for this hyperparameter tuning experiment.	○	○
You can use the policy parameter in the HyperDriveConfig class to specify a security policy.	○	○
The experiment will create a run for every possible value for the learning rate parameter between 0.05 and 0.15.	○	○

Box 1: No -
max_total_runs (50 here)
The maximum total number of runs to create. This is the upper bound; there may be fewer runs

when the sample space is smaller than this value.

Box 2: Yes -

Policy EarlyTerminationPolicy -
The early termination policy to use. If None - the default, no early termination policy will be used.

Box 3: No -
Discrete hyperparameters are specified as a choice among discrete values. choice can be:
- one or more comma-separated values
- a range object
- any arbitrary list object

Reference:
https://docs.microsoft.com/en-us/python/api/azureml-train-core/azureml.train.hyperdrive.hyperdriveconfig https://docs.microsoft.com/en-us/azure/machine-learning/how-to-tune-hyperparameters

Question #13*Topic 3*
HOTSPOT -
You are using Azure Machine Learning to train machine learning models. You need a compute target on which to remotely run the training script.
You run the following Python code:

```
from azureml.core.compute import ComputeTarget, AmlCompute
from azureml.core.compute_target import ComputeTargetException
the_cluster_name = "NewCompute"
config = AmlCompute.provisioning_configuration(vm_size= 'STANDARD_D2', max_nodes=3)
the_cluster = ComputeTarget.create(ws, the_cluster_name, config)
```

For each of the following statements, select Yes if the statement is true. Otherwise, select No.
NOTE: Each correct selection is worth one point.
Hot Area:

Answer Area

	Yes	No
The compute is created in the same region as the Machine Learning service workspace.	○	○
The compute resource created by the code is displayed as a compute cluster in Azure Machine Learning studio.	○	○
The minimum number of nodes will be zero.	○	○

Correct Answer:

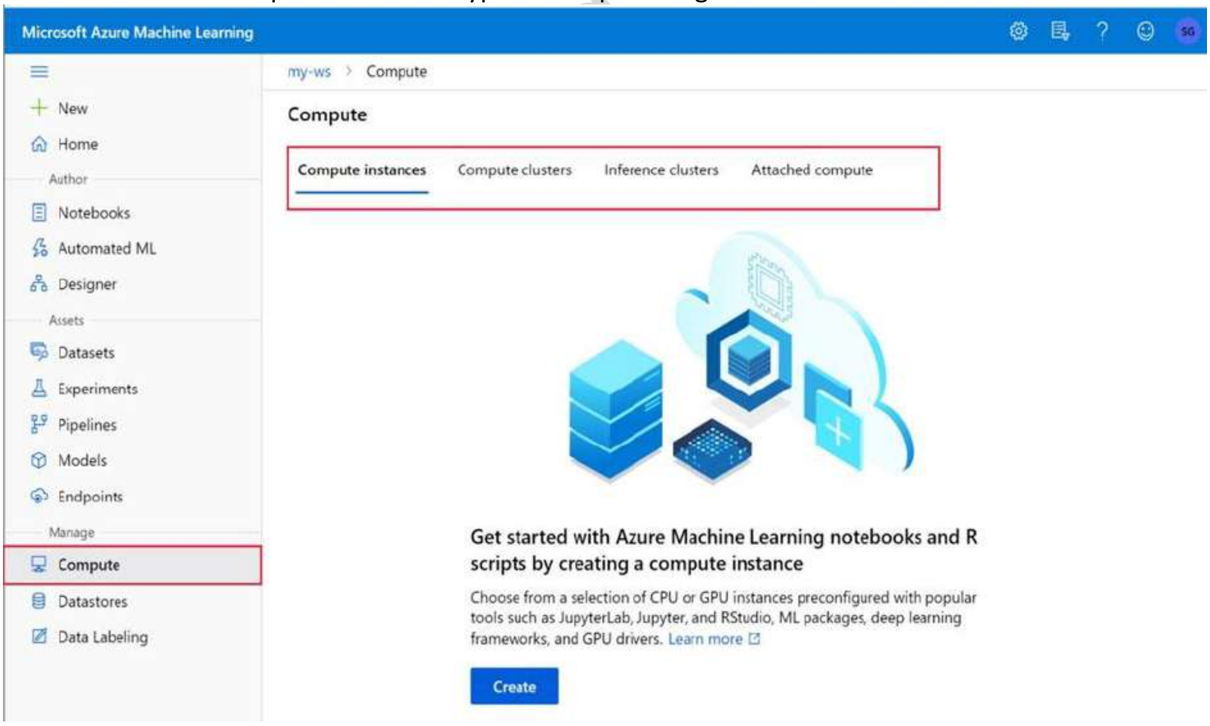
Answer Area

	Yes	No
The compute is created in the same region as the Machine Learning service workspace.	○	○
The compute resource created by the code is displayed as a compute cluster in Azure Machine Learning studio.	○	○
The minimum number of nodes will be zero.	○	○

Box 1: Yes -
The compute is created within your workspace region as a resource that can be shared with other users.

Box 2: Yes -
It is displayed as a compute cluster.

View compute targets -
1. To see all compute targets for your workspace, use the following steps:
2. Navigate to Azure Machine Learning studio.
3. Under Manage, select Compute.
4. Select tabs at the top to show each type of compute target.



Box 3: Yes -
min_nodes is not specified, so it defaults to 0.
Reference:
https://docs.microsoft.com/en-us/python/api/azureml-

core/azureml.core.compute.amlcompute.amlcomputeprovisioningconfiguration
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-create-attach-compute-studio
Question #14*Topic 3*

HOTSPOT -
You have an Azure blob container that contains a set of TSV files. The Azure blob container is registered as a datastore for an Azure Machine Learning service workspace. Each TSV file uses the same data schema.
You plan to aggregate data for all of the TSV files together and then register the aggregated data as a dataset in an Azure Machine Learning workspace by using the Azure Machine Learning SDK for Python.
You run the following code.

```
from azureml.core.workspace import Workspace
from azureml.core.datastore import Datastore
from azureml.core.dataset import Dataset
import pandas as pd
datastore_paths = (datastore, './data/*.tsv')
myDataset_1 = Dataset.File.from_files(path=datastore_paths)
myDataset_2 = Dataset.Tabular.from_delimited_files(path=datastore_paths, separator='\t')
```

For each of the following statements, select Yes if the statement is true. Otherwise, select No.
NOTE: Each correct selection is worth one point.
Hot Area:

Answer Area

	Yes	No
The myDataset_1 dataset can be converted into a pandas dataframe by using the following method: `using myDataset_1.to_pandas_dataframe()`	○	○
The myDataset_1.to_path() method returns an array of file paths for all of the TSV files in the dataset.	○	○
The myDataset_2 dataset can be converted into a pandas dataframe by using the following method: `myDataset_2.to_pandas_dataframe()`	○	○

Correct Answer:

Answer Area

	Yes	No
The myDataset_1 dataset can be converted into a pandas dataframe by using the following method: `using myDataset_1.to_pandas_dataframe()`	○	○
The myDataset_1.to_path() method returns an array of file paths for all of the TSV files in the dataset.	○	○
The myDataset_2 dataset can be converted into a pandas dataframe by using the following method: `myDataset_2.to_pandas_dataframe()`	○	○

Box 1: No -
FileDataset references single or multiple files in datastores or from public URLs. The TSV files need to be parsed.

Box 2: Yes -
to_path() gets a list of file paths for each file stream defined by the dataset.

Box 3: Yes -
TabularDataset.to_pandas_dataframe loads all records from the dataset into a pandas DataFrame. TabularDataset represents data in a tabular format created by parsing the provided file or list of files.
Note: TSV is a file extension for a tab-delimited file used with spreadsheet software. TSV stands for Tab Separated Values. TSV files are used for raw data and can be imported into and exported from spreadsheet software. TSV files are essentially text files, and the raw data can be viewed by text editors, though they are often used when moving raw data between spreadsheets.
Reference:
https://docs.microsoft.com/en-us/python/api/azureml-core/azureml.data.tabulardataset
Question #15*Topic 3*

You create a batch inference pipeline by using the Azure ML SDK. You configure the pipeline parameters by executing the following code:

```
from azureml.contrib.pipeline.steps import ParallelRunConfig
parallel_run_config = ParallelRunConfig(
      source_directory=scripts_folder,
      entry_script= “batch_pipeline.py”,
      mini_batch_size= “5”,
      error_threshold=10,
      output_action= “append_row”,
      environment=batch_env,
      compute_target=compute_target,
      logging_level= “DEBUG”,
      node_count=4)
```

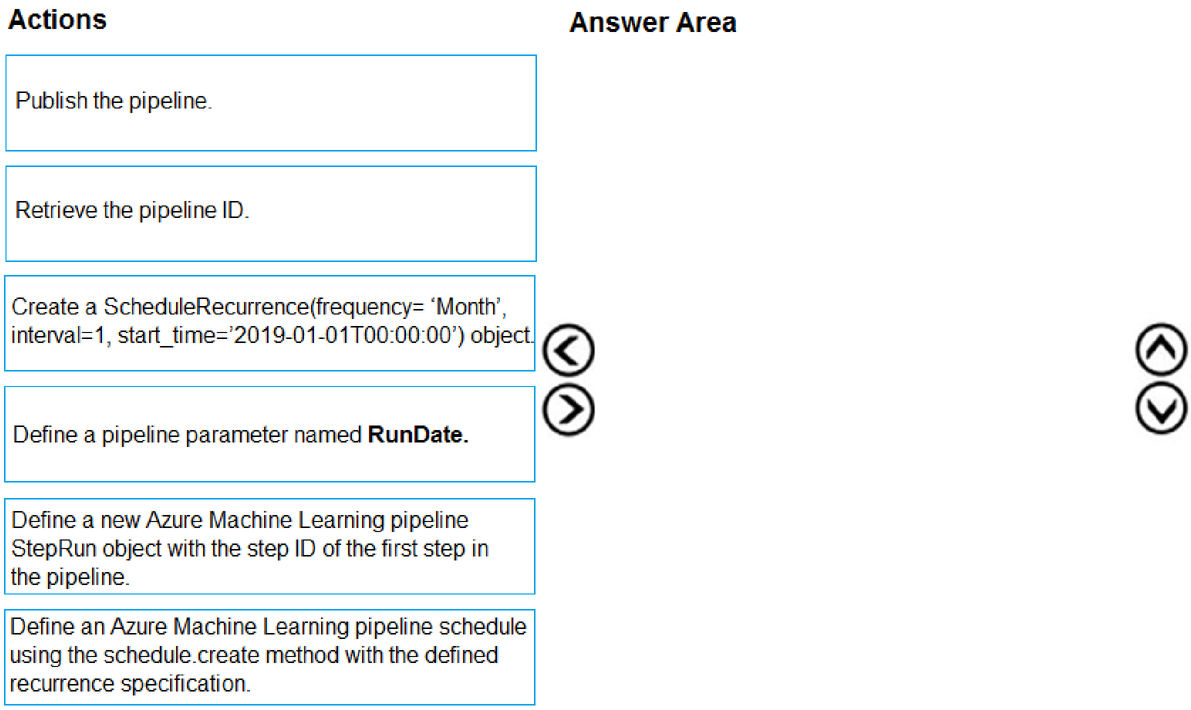
You need to obtain the output from the pipeline execution.
Where will you find the output?

- A. the digit_identification.py script
- B. the debug log
- C. the Activity Log in the Azure portal for the Machine Learning workspace
- D. the Inference Clusters tab in Machine Learning studio
- E. a file named parallel_run_step.txt located in the output folder

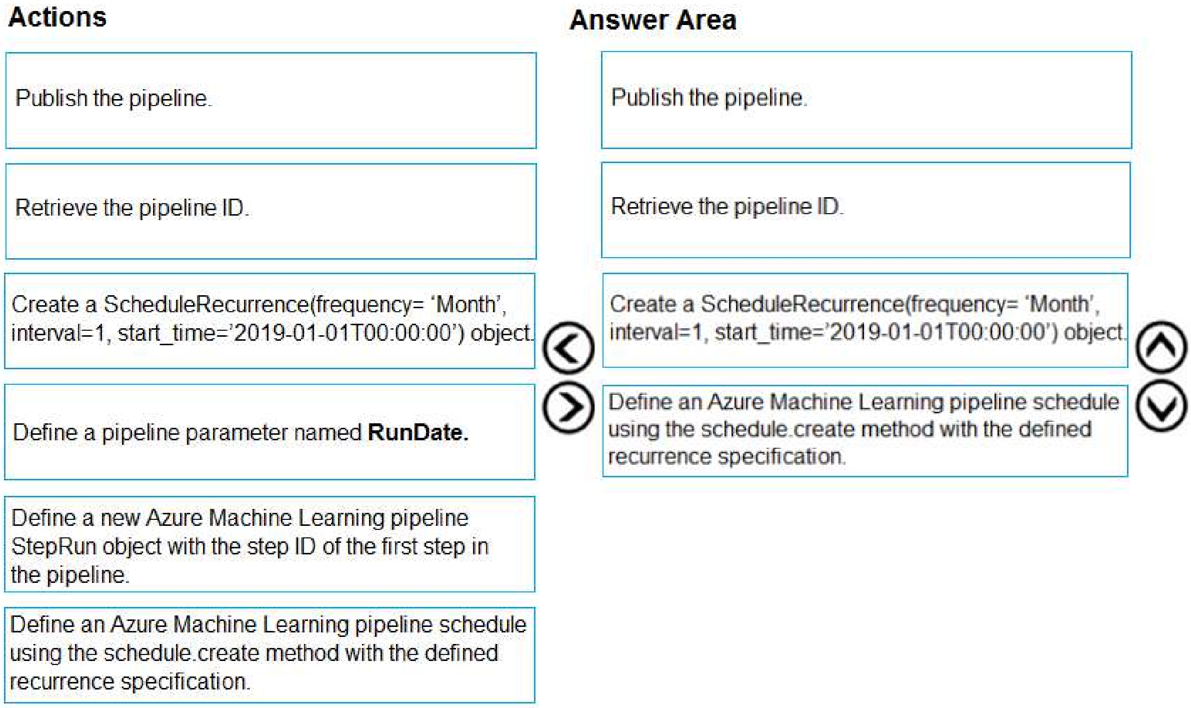
Correct Answer: *E*
output_action (str): How the output is to be organized. Currently supported values are 'append_row' and 'summary_only'.
'append_row' ג€" All values output by run() method invocations will be aggregated into one unique file named parallel_run_step.txt that is created in the output location.
'summary_only'
Reference:
https://docs.microsoft.com/en-us/python/api/azureml-contrib-pipeline-steps/azureml.contrib.pipeline.steps.parallelrunconfig

Question #16*Topic 3*

DRAG DROP -
You create a multi-class image classification deep learning model.
The model must be retrained monthly with the new image data fetched from a public web portal.
You create an Azure Machine Learning pipeline to fetch new data, standardize the size of images, and retrain the model.
You need to use the Azure Machine Learning SDK to configure the schedule for the pipeline.
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.
Select and Place:



Correct Answer:



Step 1: Publish the pipeline.
To schedule a pipeline, you'll need a reference to your workspace, the identifier of your published pipeline, and the name of the experiment in which you wish to create the schedule.
Step 2: Retrieve the pipeline ID.
Needed for the schedule.
Step 3: Create a ScheduleRecurrence..
To run a pipeline on a recurring basis, you'll create a schedule. A Schedule associates a pipeline, an experiment, and a trigger.
First create a schedule. Example: Create a Schedule that begins a run every 15 minutes: recurrence = ScheduleRecurrence(frequency="Minute", interval=15)
Step 4: Define an Azure Machine Learning pipeline schedule..
Example, continued:
recurring_schedule = Schedule.create(ws, name="MyRecurringSchedule", description="Based on time", pipeline_id=pipeline_id, experiment_name=experiment_name, recurrence=recurrence)
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-schedule-pipelines
Question #17*Topic 3*

HOTSPOT -
You create a script for training a machine learning model in Azure Machine Learning service.
You create an estimator by running the following code:

```
from azureml.core import Workspace, Datastore
from azureml.core.compute import ComputeTarget
from azureml.train.estimator import Estimator
work_space = Workspace.from_config()
data_source = work_space.get_default_datastore()
train_cluster = ComputeTarget(workspace=work_space, name= 'train-cluster')
estimator = Estimator(source_directory =
 'training-experiment',
script_params = { ' --data-folder' : data_source.as_mount(), ' --regularization':0.8},
compute_target = train_cluster,
entry_script = 'train.py',
conda_packages = ['scikit-learn'])
```

For each of the following statements, select Yes if the statement is true. Otherwise, select No.
NOTE: Each correct selection is worth one point.
Hot Area:

Answer Area

	Yes	No
The estimator will look for the files it needs to run an experiment in the training-experiment directory of the local compute environment.	○	○
The estimator will mount the local data-folder folder and make it available to the script through a parameter.	○	○
The train.py script file will be created if it does not exist.	○	○
The estimator can run Scikit-learn experiments.	○	○

Correct Answer:

Answer Area

	Yes	No
The estimator will look for the files it needs to run an experiment in the training-experiment directory of the local compute environment.	●	○
The estimator will mount the local data-folder folder and make it available to the script through a parameter.	●	○
The train.py script file will be created if it does not exist.	○	●
The estimator can run Scikit-learn experiments.	●	○

Box 1: Yes -
Parameter source_directory is a local directory containing experiment configuration and code files needed for a training job.

Box 2: Yes -

script_params is a dictionary of command-line arguments to pass to the training script specified in entry_script.

Box 3: No -

Box 4: Yes -
The conda_packages parameter is a list of strings representing conda packages to be added to the Python environment for the experiment.
Question #18*Topic 3*

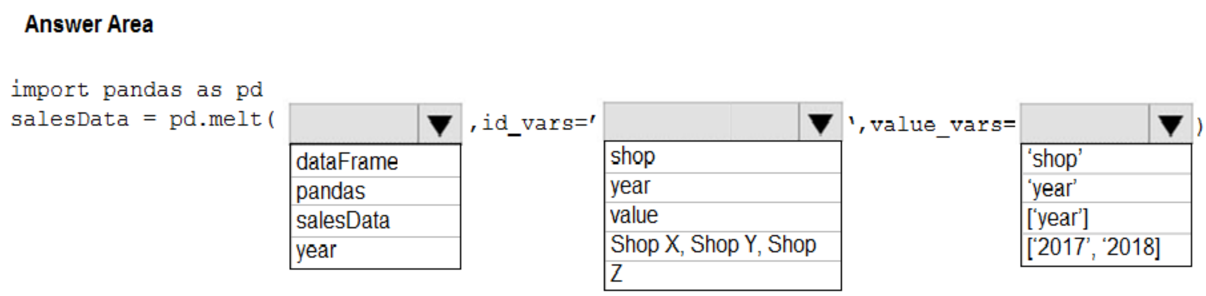
HOTSPOT -
You have a Python data frame named salesData in the following format:

	shop	2017	2018
0	Shop X	34	25
1	Shop Y	65	76
2	Shop Z	48	55

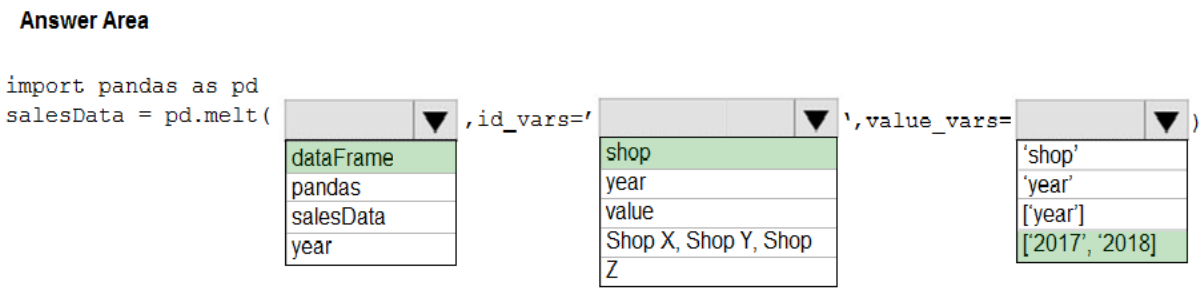
The data frame must be unpivoted to a long data format as follows:

	shop	year	value
0	Shop X	2017	34
1	Shop Y	2017	65
2	Shop Z	2017	48
3	Shop X	2018	25
4	Shop Y	2018	76
5	Shop Z	2018	55

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.
Hot Area:



Correct Answer:



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
Reference:
https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.melt.html
Question #19*Topic 3*

HOTSPOT -
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:

Name	Description
IsPlaySoccer	Values can be 1 and 0.
Gender	Values can be M or F.
PrevExamMarks	Stores values from 0 to 100
Height	Stores values in centimeters
Weight	Stores values in kilograms

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.
Hot Area:

Answer Area

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

Correct Answer:

Answer Area

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

Reference:
https://www.edureka.co/blog/classification-algorithms/
Question #20*Topic 3*

HOTSPOT -
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.

Hot Area:

Answer Area

Preprocess Text

Language

English

Remove by part of speech

False

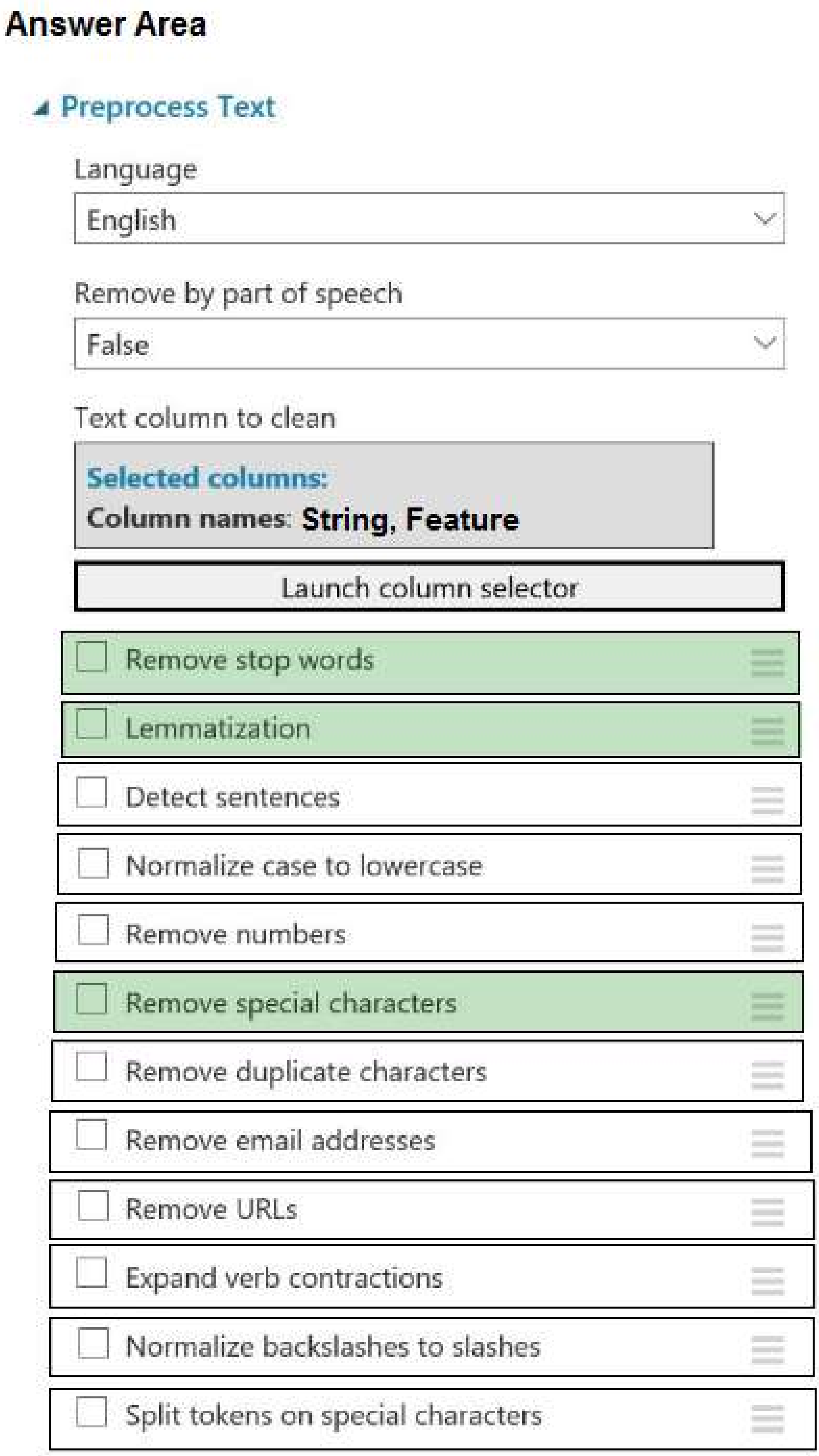
Text column to clean

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

Correct


Answer Area
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:

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.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/preprocess-text
Question #21*Topic 3*

You plan to run a script as an experiment using a Script Run Configuration. The script uses modules from the scipy library as well as several Python packages that are not typically installed in a default conda environment.
You plan to run the experiment on your local workstation for small datasets and scale out the experiment by running it on more powerful remote compute clusters for larger datasets.
You need to ensure that the experiment runs successfully on local and remote compute with the least administrative effort.
What should you do?

- A. Do not specify an environment in the run configuration for the experiment. Run the experiment by using the default environment.
- B. Create a virtual machine (VM) with the required Python configuration and attach the VM as a compute target. Use this compute target for all experiment runs.
- C. Create and register an Environment that includes the required packages. Use this Environment for all experiment runs.
- D. Create a config.yaml file defining the conda packages that are required and save the file in the experiment folder.
- E. Always run the experiment with an Estimator by using the default packages.

Correct Answer: *C*
If you have an existing Conda environment on your local computer, then you can use the service to create an environment object. By using this strategy, you can reuse your local interactive environment on remote runs.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-use-environments

Question #22*Topic 3*

You write a Python script that processes data in a comma-separated values (CSV) file.
You plan to run this script as an Azure Machine Learning experiment.
The script loads the data and determines the number of rows it contains using the following code:

```
from azureml.core import Run
import pandas as pd

run = Run.get_context()
data = pd.read_csv('./data.csv')
rows = (len(data))
# record row_count metric here
...
```

You need to record the row count as a metric named row_count that can be returned using the get_metrics method of the Run object after the experiment run completes.
Which code should you use?

- A. run.upload_file(T3 row_count', './data.csv')
- B. run.log('row_count', rows)
- C. run.tag('row_count', rows)
- D. run.log_table('row_count', rows)
- E. run.log_row('row_count', rows)

Correct Answer: *B*
Log a numerical or string value to the run with the given name using log(name, value, description='').
Logging a metric to a run causes that metric to be stored in the run record in the experiment. You can log the same metric multiple times within a run, the result being considered a vector of that metric.
Example: run.log("accuracy", 0.95)
Incorrect Answers:
E: Using log_row(name, description=None, **kwargs) creates a metric with multiple columns as described in kwargs. Each named parameter generates a column with the value specified. log_row can be called once to log an arbitrary tuple, or multiple times in a loop to generate a complete table.
Example: run.log_row("Y over X", x=1, y=0.4)
Reference:
https://docs.microsoft.com/en-us/python/api/azureml-core/azureml.core.run

Question #23*Topic 3*

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 Synthetic Minority Oversampling Technique (SMOTE) sampling mode.
Does the solution meet the goal?

- A. Yes

- B. No

Correct Answer: *A*
SMOTE is used 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.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/smote

Question #24*Topic 3*

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 Stratified split for the sampling mode.
Does the solution meet the goal?

- A. Yes
- B. No

Correct Answer: *B*
Instead use the Synthetic Minority Oversampling Technique (SMOTE) sampling mode.
Note: SMOTE is used 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.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/smote

Question #25*Topic 3*

You are creating a machine learning model.
You need to identify outliers in the data.
Which two visualizations can you use? Each correct answer presents a complete solution.
NOTE: Each correct selection is worth one point.

- A. Venn diagram
- B. Box plot
- C. ROC curve
- D. Random forest diagram
- E. Scatter plot

Correct Answer: *BE*
The box-plot algorithm can be used to display outliers.
One other way to quickly identify Outliers visually is to create scatter plots.
Reference:

https://blogs.msdn.microsoft.com/azuredev/2017/05/27/data-cleansing-tools-in-azure-machine-learning/

Question #26*Topic 3*

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. Violin plot
- B. Gradient descent
- C. Box plot
- D. Binary classification confusion matrix

Correct Answer: *D*
Incorrect Answers:
A: A violin plot is a visual that traditionally combines a box plot and a kernel density plot.
B: Gradient descent is a first-order iterative optimization algorithm for finding the minimum of a function. To find a local minimum of a function using gradient descent, one takes steps proportional to the negative of the gradient (or approximate gradient) of the function at the current point.
C: A box plot lets you see basic distribution information about your data, such as median, mean, range and quartiles but doesn't show you how your data looks throughout its range.
Reference:
https://machinelearningknowledge.ai/confusion-matrix-and-performance-metrics-machine-learning/

Question #27*Topic 3*

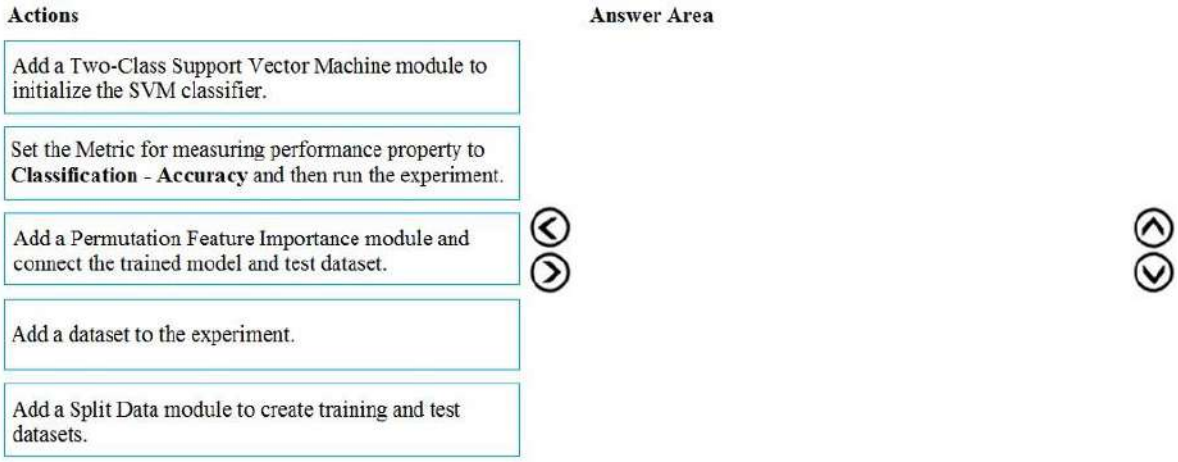
You create a multi-class image classification deep learning model that uses the PyTorch deep learning framework.
You must configure Azure Machine Learning Hyperdrive to optimize the hyperparameters for the classification model.
You need to define a primary metric to determine the hyperparameter values that result in the model with the best accuracy score.
Which three actions must you perform? Each correct answer presents part of the solution.
NOTE: Each correct selection is worth one point.

- A. Set the primary_metric_goal of the estimator used to run the bird_classifier_train.py script to maximize.
- B. Add code to the bird_classifier_train.py script to calculate the validation loss of the model and log it as a float value with the key loss.
- C. Set the primary_metric_goal of the estimator used to run the bird_classifier_train.py script to minimize.
- D. Set the primary_metric_name of the estimator used to run the bird_classifier_train.py script to accuracy.
- E. Set the primary_metric_name of the estimator used to run the bird_classifier_train.py script to loss.
- F. Add code to the bird_classifier_train.py script to calculate the validation accuracy of the model and log it as a float value with the key accuracy.

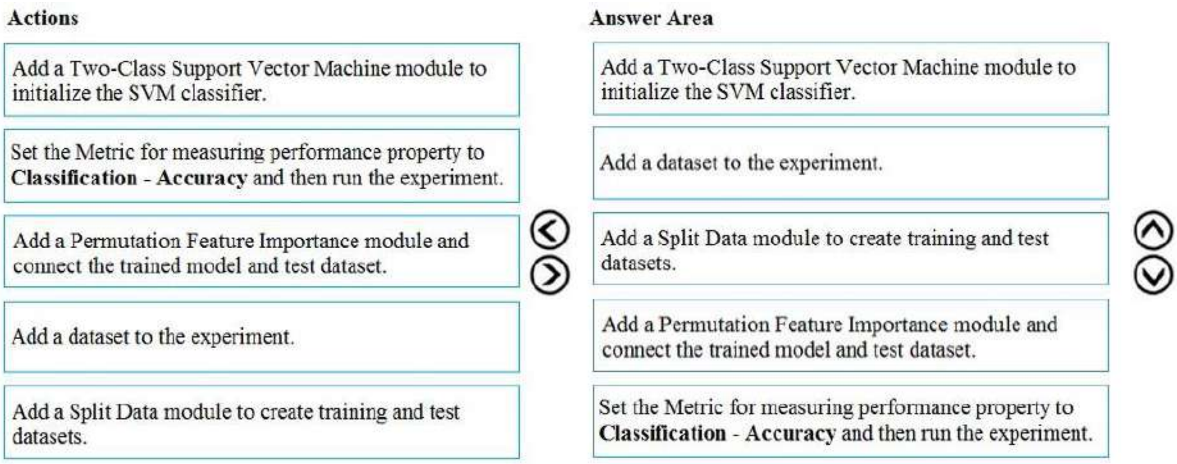
Correct Answer: *ADF*
AD:
primary_metric_name="accuracy",
primary_metric_goal=PrimaryMetricGoal.MAXIMIZE
Optimize the runs to maximize "accuracy". Make sure to log this value in your training script.
Note:
primary_metric_name: The name of the primary metric to optimize. The name of the primary metric needs to exactly match the name of the metric logged by the training script. primary_metric_goal: It can be either PrimaryMetricGoal.MAXIMIZE or PrimaryMetricGoal.MINIMIZE and determines whether the primary metric will be maximized or minimized when evaluating the runs.
F: The training script calculates the val_accuracy and logs it as "accuracy", which is used as the primary metric.

Question #28*Topic 3*

DRAG DROP -
You have a dataset that contains over 150 features. You use the dataset to train a Support Vector Machine (SVM) binary classifier.
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 all actions from the list of actions to the answer area and arrange them in the correct order.
Select and Place:



Correct Answer:



Step 1: Add a Two-Class Support Vector Machine module to initialize the SVM classifier.
Step 2: Add a dataset to the experiment
Step 3: Add a Split Data module to create training and test dataset.
To generate a set of feature scores requires that you have an already trained model, as well as a test dataset.
Step 4: Add a Permutation Feature Importance module and connect to the trained model and test dataset.
Step 5: Set the Metric for measuring performance property to Classification - Accuracy and then run the experiment.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/two-class-support-vector-machine https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/permutation-feature-importance
Question #29*Topic 3*

HOTSPOT -
You are using the Hyperdrive feature in Azure Machine Learning to train a model.
You configure the Hyperdrive experiment by running the following code:

```
from azureml.train.hyperdrive import RandomParameterSampling
param_sampling = RandomParameterSampling( {
        "learning_rate": normal(10, 3),
        "keep_probability": uniform(0.05, 0.1),
        "batch_size": choice(16, 32, 64, 128)
        "number_of_hidden_layers": choice(range(3,5))
    }
)
```

For each of the following statements, select Yes if the statement is true. Otherwise, select No.
NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

	Yes	No
By defining sampling in this manner, every possible combination of the parameters will be tested.	○	○
Random values of the learning_rate parameter will be selected from a normal distribution with a mean of 10 and a standard deviation of 3.	○	○
The keep_probability parameter value will always be either **0.05** or **0.1**.	○	○
Random values for the number_of_hidden_layers parameter will be selected from a normal distribution with a mean of 3 and a standard deviation of 5.	○	○

Correct Answer:

Answer Area

	Yes	No
By defining sampling in this manner, every possible combination of the parameters will be tested.	○	○
Random values of the learning_rate parameter will be selected from a normal distribution with a mean of 10 and a standard deviation of 3.	○	○
The keep_probability parameter value will always be either **0.05** or **0.1**.	○	○
Random values for the number_of_hidden_layers parameter will be selected from a normal distribution with a mean of 3 and a standard deviation of 5.	○	○

Box 1: Yes -
In random sampling, hyperparameter values are randomly selected from the defined search space. Random sampling allows the search space to include both discrete and continuous hyperparameters.

Box 2: Yes -
learning_rate has a normal distribution with mean value 10 and a standard deviation of 3.

Box 3: No -
keep_probability has a uniform distribution with a minimum value of 0.05 and a maximum value of 0.1.

Box 4: No -
number_of_hidden_layers takes on one of the values [3, 4, 5].
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-tune-hyperparameters
Question #30*Topic 3*

You are performing a filter-based feature selection for a dataset to build a multi-class classifier 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. Kendall correlation
- B. Spearman correlation

- C. Chi-squared
- D. Pearson correlation

Correct Answer: *C*
Pearson's correlation statistic, or Pearson's correlation coefficient, is also known in statistical models as the r value. For any two variables, it returns a value that indicates the strength of the correlation Pearson's correlation coefficient is the test statistics that measures the statistical relationship, or association, between two continuous variables. It is known as the best method of measuring the association between variables of interest because it is based on the method of covariance. It gives information about the magnitude of the association, or correlation, as well as the direction of the relationship.
Incorrect Answers:
C: The two-way chi-squared test is a statistical method that measures how close expected values are to actual results.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/filter-based-feature-selection https://www.statisticssolutions.com/pearsons-correlation-coefficient/

Question #31*Topic 3*

HOTSPOT -
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.

Hot Area:

Answer Area

Description	Error type
A person has a disease. The model classifies the case as having a disease.	True Positives True Negatives False Positives False Negatives
A person does not have a disease. The model classifies the case as having no disease.	True Positives True Negatives False Positives False Negatives
A person does not have a disease. The model classifies the case as having a disease.	True Positives True Negatives False Positives False Negatives
A person has a disease. The model classifies the case as having no disease.	True Positives True Negatives False Positives False Negatives

Correct Answer:

Answer Area

Description	Error type
A person has a disease. The model classifies the case as having a disease.	True Positives True Negatives False Positives False Negatives
A person does not have a disease. The model classifies the case as having no disease.	True Positives True Negatives False Positives False Negatives
A person does not have a disease. The model classifies the case as having a disease.	True Positives True Negatives False Positives False Negatives
A person has a disease. The model classifies the case as having no disease.	True Positives True Negatives False Positives False Negatives

Box 1: True Positive -
A true positive is an outcome where the model correctly predicts the positive class

Box 2: True Negative -
A true negative is an outcome where the model correctly predicts the negative class.

Box 3: False Positive -
A false positive is an outcome where the model incorrectly predicts the positive class.

Box 4: False Negative -
A false negative is an outcome where the model incorrectly predicts the negative class.
Note: Let's make the following definitions:
"Wolf" is a positive class.
"No wolf" is a negative class.
We can summarize our "wolf-prediction" model using a 2x2 confusion matrix that depicts all four possible outcomes:

Reference:
https://developers.google.com/machine-learning/crash-course/classification/true-false-positive-negative
Question #32*Topic 3*

HOTSPOT -
You are using the Azure Machine Learning Service to automate hyperparameter exploration of your neural network classification model.
You must define the hyperparameter space to automatically tune hyperparameters using random sampling according to following requirements:
- The 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 param_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.
Hot Area:

Answer Area

```
from azureml.train.hyperdrive import RandomParameterSampling
param_sampling = RandomParameterSampling( {

"learning_rate" :  [ ▼ ] ,
    uniform(10,3)
    normal(10,3)
    choice(10,3)
    Loguniform(10,3)

"batch_size":  [ ▼ ] ,
    choice(16,32,64)
    choice(range(16,64))
    normal(16,32,64)
    normal(range(16,64))

"keep_probability" :  [ ▼ ]
    choice(range(0.05, 0.1))
    uniform(0.05, 0.1)
    normal(0.05, 0.1)
    lognormal(0.05, 0.1)

    }
)
```

Correct Answer:

Answer Area

```
from azureml.train.hyperdrive import RandomParameterSampling
param_sampling = RandomParameterSampling( {
"learning_rate" :
```

uniform(10,3)
normal(10,3)
choice(10,3)
Loguniform(10,3)

```
"batch_size":
```

choice(16,32,64)
choice(range(16,64))
normal(16,32,64)
normal(range(16,64))

```
"keep_probability" :
```

choice(range(0.05, 0.1))
uniform(0.05, 0.1)
normal(0.05, 0.1)
lognormal(0.05, 0.1)

```
    }
)
```

Box 1: normal(10,3)
Box 2: choice(16, 32, 64)
Box 3: uniform(0.05, 0.1)
In random sampling, hyperparameter values are randomly selected from the defined search space.
Random sampling allows the search space to include both discrete and continuous hyperparameters.
Example:
from azureml.train.hyperdrive import RandomParameterSampling
param_sampling = RandomParameterSampling({
"learning_rate": normal(10, 3),
"keep_probability": uniform(0.05, 0.1),
"batch_size": choice(16, 32, 64)
}
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/service/how-to-tune-hyperparameters
Question #33*Topic 3*

You plan to use automated machine learning to train a regression model. You have data that has features which have missing values, and categorical features with few distinct values.
You need to configure automated machine learning to automatically impute missing values and encode categorical features as part of the training task.
Which parameter and value pair should you use in the AutoMLConfig class?

- A. featurization = 'auto'
- B. enable_voting_ensemble = True
- C. task = 'classification'
- D. exclude_nan_labels = True
- E. enable_tf = True

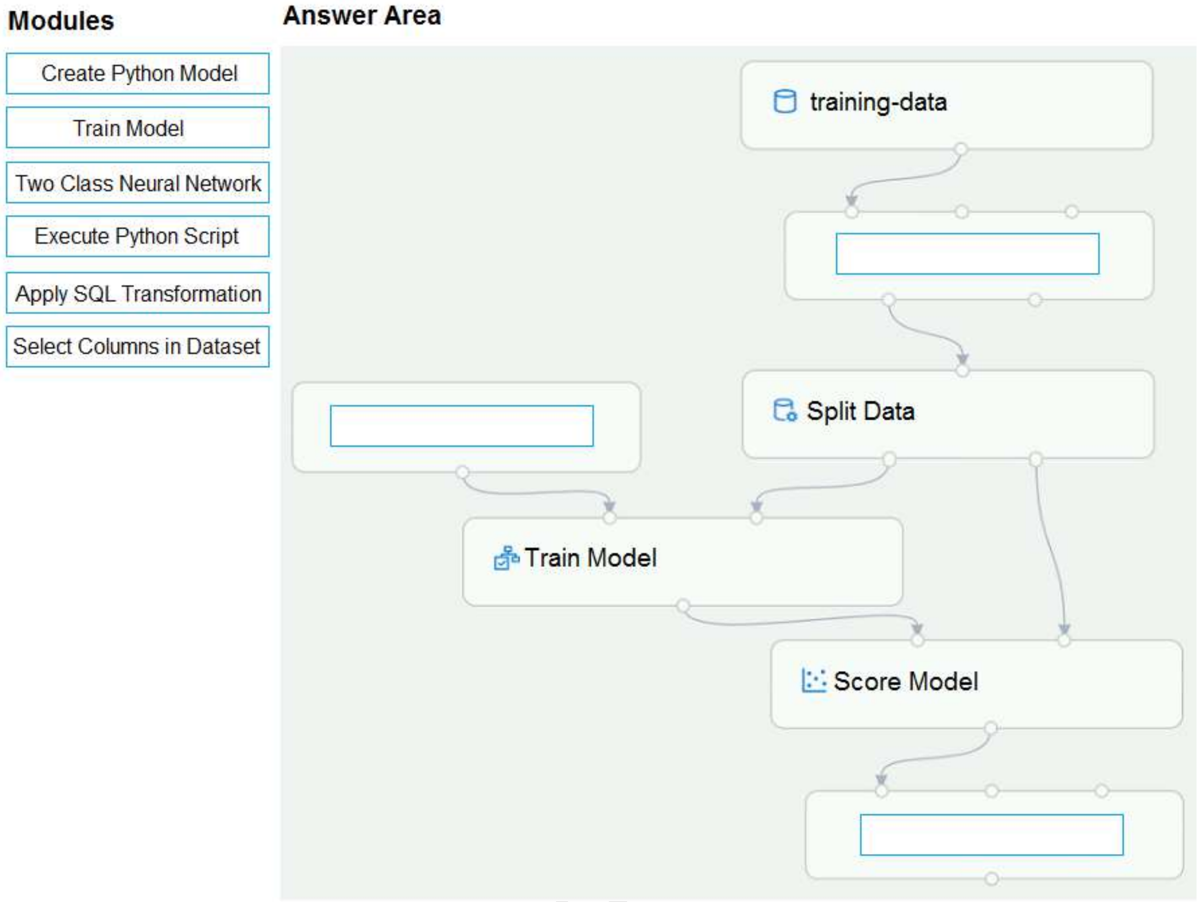
Correct Answer: *A*
Featurization str or FeaturizationConfig
Values: 'auto' / 'off' / FeaturizationConfig
Indicator for whether featurization step should be done automatically or not, or whether customized featurization should be used.
Column type is automatically detected. Based on the detected column type preprocessing/featurization is done as follows:
Categorical: Target encoding, one hot encoding, drop high cardinality categories, impute missing values.
Numeric: Impute missing values, cluster distance, weight of evidence.
DateTime: Several features such as day, seconds, minutes, hours etc.
Text: Bag of words, pre-trained Word embedding, text target encoding.
Reference:
https://docs.microsoft.com/en-us/python/api/azureml-train-automl-client/azureml.train.automl.automlconfig.automlconfig

Question #34*Topic 3*

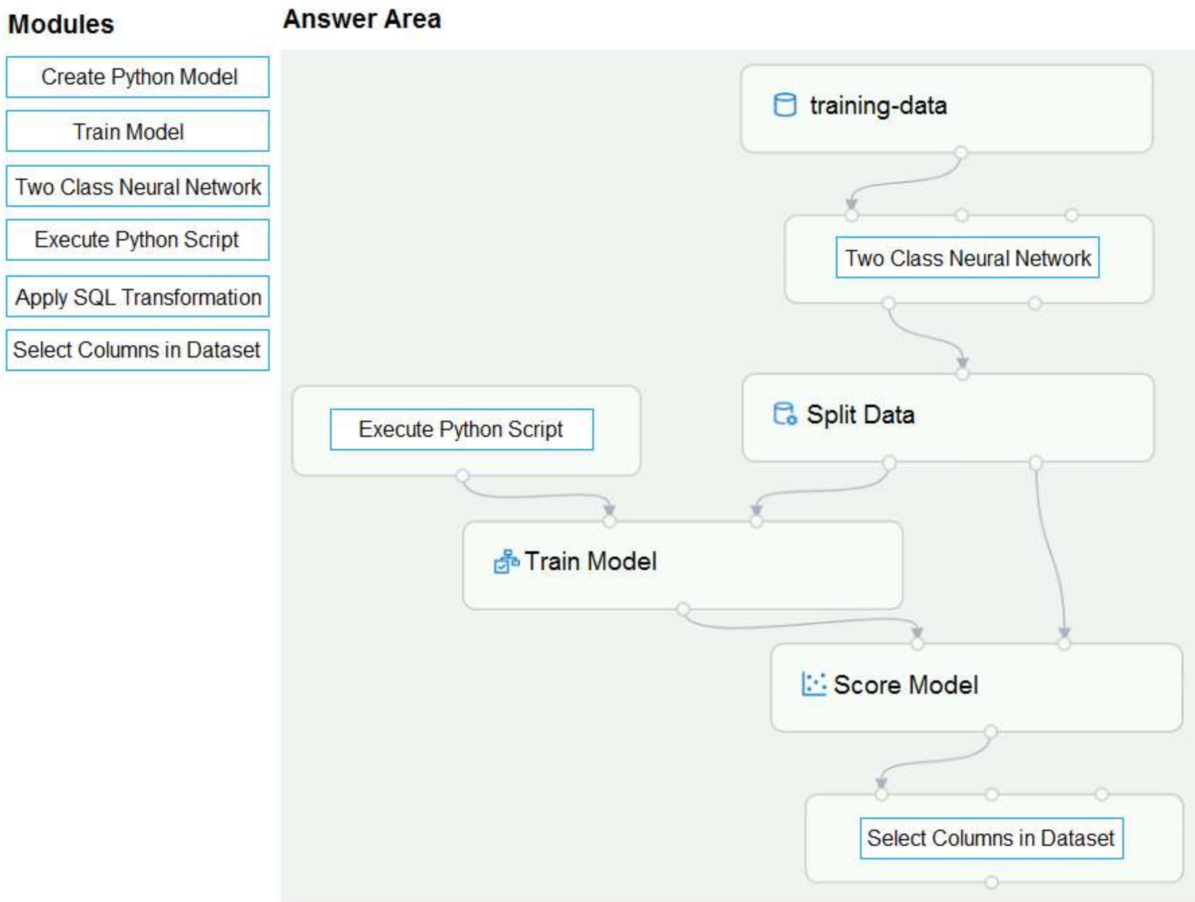
DRAG DROP -
You create a training pipeline using the Azure Machine Learning designer. You upload a CSV file that contains the data from which you want to train your model.
You need to use the designer to create a pipeline that includes steps to perform the following tasks:
- Select the training features using the pandas filter method.
- Train a model based on the naive_bayes.GaussianNB algorithm.
- Return only the Scored Labels column by using the query
- SELECT [Scored Labels] FROM t1;

Which modules should you use? To answer, drag the appropriate modules to the appropriate locations. Each module name 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.

Select and Place:



Correct Answer:



Box 1: Two-Class Neural Network -
The Two-Class Neural Network creates a binary classifier using a neural network algorithm.
Train a model based on the naive_bayes.GaussianNB algorithm.

Box 2: Execute python script -
Select the training features using the pandas filter method
Box 3: Select Columns in DataSet
Return only the Scored Labels column by using the query SELECT [Scored Labels] FROM t1;
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/two-class-neural-network
Question #35*Topic 3*

You are building a regression model for 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? Each 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 must be whole numbers.
- C. The label data must be non-discrete.
- D. The label data must be a positive value.
- E. The label data can be positive or negative.

Correct Answer: *BD*
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.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/poisson-regression

Question #36*Topic 3*

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 Principal Components Analysis (PCA) sampling mode.
Does the solution meet the goal?

- A. Yes
- B. No

Correct Answer: *B*
Instead use the Synthetic Minority Oversampling Technique (SMOTE) sampling mode.
Note: SMOTE is used 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.
Incorrect Answers:
The Principal Component Analysis module in Azure Machine Learning Studio (classic) is used to reduce the dimensionality of your training data. The module analyzes your data and creates a reduced feature set that captures all the information contained in the dataset, but in a smaller number of features.
Reference:
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/principal-component-analysis

Question #37*Topic 3*

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. Filter Based Feature Selection
- C. Execute Python Script
- D. Latent Dirichlet Allocation

Correct Answer: *C*
Typical metadata changes might include marking columns as features.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/edit-metadata

Question #38*Topic 3*

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. violin plot
- B. Gradient descent
- C. Scatter plot
- D. Receiver Operating Characteristic (ROC) curve

Correct Answer: *D*
Receiver operating characteristic (or ROC) is a plot of the correctly classified labels vs. the incorrectly classified labels for a particular model.
Incorrect Answers:
A: A violin plot is a visual that traditionally combines a box plot and a kernel density plot.
B: Gradient descent is a first-order iterative optimization algorithm for finding the minimum of a function. To find a local minimum of a function using gradient descent, one takes steps proportional to the negative of the gradient (or approximate gradient) of the function at the current point.
C: A scatter plot graphs the actual values in your data against the values predicted by the model. The scatter plot displays the actual values along the X-axis, and displays the predicted values along the Y-axis. It also displays a line that illustrates the perfect prediction, where the predicted value exactly matches the actual value.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-understand-automated-ml#confusion-matrix

Question #39*Topic 3*

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=1
- B. k=10
- C. k=0.5
- D. k=0.9

Correct Answer: *B*
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.

Question #40*Topic 3*

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

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

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

Hot Area:

Answer Area

```
from sklearn.decomposition import PCA
pca = [▼]
```

- PCA()
- PCA(n_components = 150)
- PCA(n_components = 10)
- PCA(n_components = 10000)

```
X_train= [▼] .fit_transform(X_train)
```

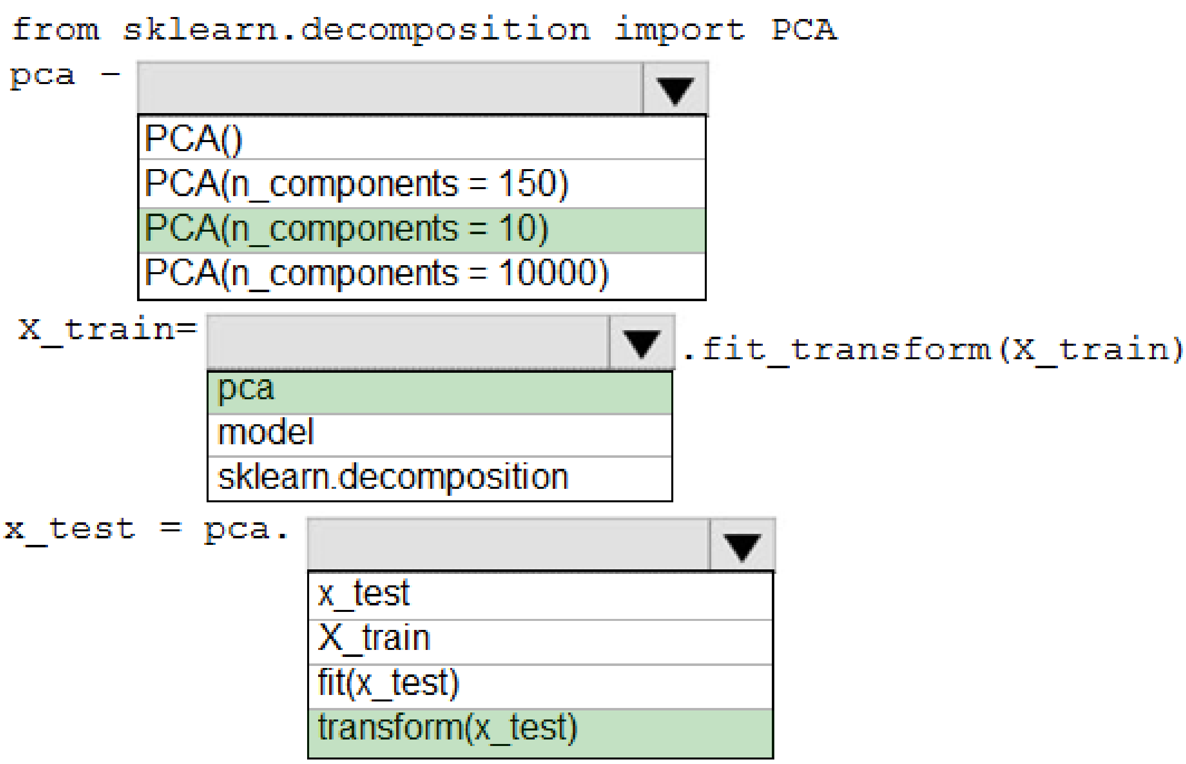
- pca
- model
- sklearn.decomposition

```
x_test = pca. [▼]
```

- x_test
- X_train
- fit(x_test)
- transform(x_test)

Correct Answer:

Answer Area



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.
Reference:
https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html
Question #41*Topic 3*

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

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

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

NOTE: Each correct selection is worth one point.
Hot Area:

Answer Area

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

- -0.106276
- 0.149676
- 0.859122
- 1

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

- a positive linear relationship
- a negative linear relationship
- no linear relationship

Correct Answer:

Answer Area

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

- -0.106276
- 0.149676
- 0.859122
- 1

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

- a positive linear relationship
- a negative linear relationship
- no linear relationship

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.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/compute-linear-correlation
Question #42*Topic 3*

DRAG DROP -
You plan to explore demographic data for home ownership in various cities. The data is in a CSV file with the following format: age,city,income,home_owner
21,Chicago,50000,0
35,Seattle,120000,1
23,Seattle,65000,0
45,Seattle,130000,1

18,Chicago,48000,0

You need to run an experiment in your Azure Machine Learning workspace to explore the data and log the results. The experiment must log the following information:

- the number of observations in the dataset
- a box plot of income by home_owner
- a dictionary containing the city names and the average income for each city

You need to use the appropriate logging methods of the experiment's run object to log the required information.

How should you complete the code? To answer, drag the appropriate code segments to the correct locations. Each code segment 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.

Select and Place:

Code segments

- log
- log_list
- log_row
- log_table
- log_image

Answer Area

```
from azureml.core import Experiment, Run
import pandas as pd
import matplotlib.pyplot as plt
# Create an Azure ML experiment in workspace
experiment = Experiment(workspace = ws, name = "demo-experiment")
# Start logging data from the experiment
run = experiment.start_logging()
# load the dataset
data = pd.read_csv('research/demographics.csv')
# Log the number of observations
row_count = (len(data))
run. Segment ("observations", row_count)
# Log box plot for income by home_owner
fig = plt.figure(figsize=(9, 6))
ax = fig.gca()
data.boxplot(column = 'income', by = "home_owner", ax = ax)
ax.set_title('income by home_owner')
ax.set_ylabel('income')
run. Segment (name = 'income_by_home_owner', plot = fig)
# Create a dataframe of mean income per city
mean_inc_df = data.groupby('city')['income'].agg(np.mean).to_frame().reset_index()
# Convert to a dictionary
mean_inc_dict = mean_inc_df.to_dict('dict')
# Log city names and average income dictionary
run. Segment (name="mean_income_by_city", value= mean_inc_dict)
# Complete tracking and get link to details
run.complete()
```

Correct Answer:

Code segments

- log
- log_list
- log_row
- log_table
- log_image

Answer Area

```
from azureml.core import Experiment, Run
import pandas as pd
import matplotlib.pyplot as plt
# Create an Azure ML experiment in workspace
experiment = Experiment(workspace = ws, name = "demo-experiment")
# Start logging data from the experiment
run = experiment.start_logging()
# load the dataset
data = pd.read_csv('research/demographics.csv')
# Log the number of observations
row_count = (len(data))
run. log ("observations", row_count)
# Log box plot for income by home_owner
fig = plt.figure(figsize=(9, 6))
ax = fig.gca()
data.boxplot(column = 'income', by = "home_owner", ax = ax)
ax.set_title('income by home_owner')
ax.set_ylabel('income')
run. log_image (name = 'income_by_home_owner', plot = fig)
# Create a dataframe of mean income per city
mean_inc_df = data.groupby('city')['income'].agg(np.mean).to_frame().reset_index()
# Convert to a dictionary
mean_inc_dict = mean_inc_df.to_dict('dict')
# Log city names and average income dictionary
run. log_table (name="mean_income_by_city", value= mean_inc_dict)
# Complete tracking and get link to details
run.complete()
```

Box 1: log -
The number of observations in the dataset.
run.log(name, value, description='')
Scalar values: Log a numerical or string value to the run with the given name. Logging a metric to a run causes that metric to be stored in the run record in the experiment. You can log the same metric multiple times within a run, the result being considered a vector of that metric.
Example: run.log("accuracy", 0.95)

Box 2: log_image -
A box plot of income by home_owner.
log_image Log an image to the run record. Use log_image to log a .PNG image file or a matplotlib plot to the run. These images will be visible and comparable in the run record.
Example: run.log_image("ROC", plot=plt)

Box 3: log_table -
A dictionary containing the city names and the average income for each city. log_table: Log a dictionary object to the run with the given name.
Question #43*Topic 3*

You use the Azure Machine Learning service to create a tabular dataset named training_data. You plan to use this dataset in a training script.
You create a variable that references the dataset using the following code: training_ds = workspace.datasets.get("training_data")
You define an estimator to run the script.
You need to set the correct property of the estimator to ensure that your script can access the training_data dataset.
Which property should you set?

- A. environment_definition = {"training_data":training_ds}
- B. inputs = [training_ds.as_named_input('training_ds')]
- C. script_params = {"--training_ds":training_ds}

- D. source_directory = training_ds

Correct Answer: *B*
Example:
# Get the training dataset
diabetes_ds = ws.datasets.get("Diabetes Dataset")
# Create an estimator that uses the remote compute
hyper_estimator = SKLearn(source_directory=experiment_folder,
inputs=[diabetes_ds.as_named_input('diabetes')], # Pass the dataset as an input compute_target = cpu_cluster, conda_packages=['pandas','ipykernel','matplotlib'], pip_packages=['azureml-sdk','argparse','pyarrow'], entry_script='diabetes_training.py')
Reference:
https://notebooks.azure.com/GraemeMalcolm/projects/azureml-primers/html/04%20-%20Optimizing%20Model%20Training.ipynb

Question #44*Topic 3*

You register a file dataset named csv_folder that references a folder. The folder includes multiple comma-separated values (CSV) files in an Azure storage blob container.
You plan to use the following code to run a script that loads data from the file dataset. You create and instantiate the following variables:

Variable	Description
remote_cluster	References the Azure Machine Learning compute cluster
ws	References the Azure Machine Learning workspace

You have the following code:

```
from azureml.train.estimator import Estimator
file_dataset = ws.datasets.get('csv_folder')
estimator = Estimator(source_directory=script_folder,

compute_target = remote_cluster,
entry_script ='script.py')
run = experiment.submit(config=estimator)
run.wait_for_completion(show_output=True)
```

You need to pass the dataset to ensure that the script can read the files it references.
Which code segment should you insert to replace the code comment?

- A. inputs=[file_dataset.as_named_input('training_files')],
- B. inputs=[file_dataset.as_named_input('training_files').as_mount()],
- C. inputs=[file_dataset.as_named_input('training_files').to_pandas_dataframe()],
- D. script_params={'--training_files': file_dataset},

Correct Answer: *B*
Example:
from azureml.train.estimator import Estimator
script_params = {
# to mount files referenced by mnist dataset
'--data-folder': mnist_file_dataset.as_named_input('mnist_opendataset').as_mount(),

'--regularization': 0.5
}
est = Estimator(source_directory=script_folder,
script_params=script_params,
compute_target=compute_target,
environment_definition=env,
entry_script='train.py')
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/tutorial-train-models-with-aml

Question #45*Topic 3*

You are creating a new Azure Machine Learning pipeline using the designer.
The pipeline must train a model using data in a comma-separated values (CSV) file that is published on a website. You have not created a dataset for this file.
You need to ingest the data from the CSV file into the designer pipeline using the minimal administrative effort.
Which module should you add to the pipeline in Designer?

- A. Convert to CSV
- B. Enter Data Manually
- C. Import Data
- D. Dataset

Correct Answer: *C*
The preferred way to provide data to a pipeline is a Dataset object. The Dataset object points to data that lives in or is accessible from a datastore or at a Web
URL. The Dataset class is abstract, so you will create an instance of either a FileDataset (referring to one or more files) or a TabularDataset that's created by from one or more files with delimited columns of data.
Example:
from azureml.core import Dataset
iris_tabular_dataset = Dataset.Tabular.from_delimited_files([(def_blob_store, 'train-dataset/iris.csv')])
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-create-your-first-pipeline

Question #46*Topic 3*

You define a datastore named ml-data for an Azure Storage blob container. In the container, you have a folder named train that contains a file named data.csv.
You plan to use the file to train a model by using the Azure Machine Learning SDK.
You plan to train the model by using the Azure Machine Learning SDK to run an experiment on local compute.
You define a DataReference object by running the following code:

```
from azureml.core import Workspace, Datastore, Environment
from azureml.train.estimator import Estimator
ws = Workspace.from_config()
ml_data = = Datastore.get(ws, datastore_name='ml-data')
data_ref = ml_data.path('train').as_download(path_on_compute='train_data')
estimator = Estimator(source_directory='experiment_folder',
  script_params={'--data-folder': data_ref},
  compute_target = 'local',
  entry_script='training.py')
run = experiment.submit(config=estimator)
run.wait_for_completion(show_output=True)
```

You need to load the training data.
Which code segment should you use?
A.

```
import os
import argparse
import pandas as pd

parser = argparse.ArgumentParser()
parser.add_argument('--data-folder', type=str, dest='data_folder')
data_folder = args.data_folder
data = pd.read_csv(os.path.join(data_folder,'ml-data','train_data','data.csv'))
```

B.

```
import os
import argparse
import pandas as pd

parser = argparse.ArgumentParser()
parser.add_argument('--data-folder', type=str, dest='data_folder')
data_folder = args.data_folder
data = pd.read_csv(os.path.join(data_folder,'train','data.csv'))
```

C.

```
import pandas as pd

data = pd.read_csv('./data.csv')
```

D.

```
import os
import argparse
import pandas as pd

parser = argparse.ArgumentParser()
parser.add_argument('--data-folder', type=str, dest='data_folder')
data_folder = args.data_folder
data = pd.read_csv(os.path.join('ml_data', data_folder,'data.csv'))
```

E.

```
import os
import argparse
import pandas as pd

parser = argparse.ArgumentParser()
parser.add_argument('--data-folder', type=str, dest='data_folder')
data_folder = args.data_folder
data = pd.read_csv(os.path.join(data_folder,'data.csv'))
```

Correct Answer: *E*
Example:
data_folder = args.data_folder
# Load Train and Test data
train_data = pd.read_csv(os.path.join(data_folder, 'data.csv'))

Reference:
https://www.element61.be/en/resource/azure-machine-learning-services-complete-toolbox-ai
Question #47*Topic 3*

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 create an Azure Machine Learning service datastore in a workspace. The datastore contains the following files:
- /data/2018/Q1.csv
- /data/2018/Q2.csv
- /data/2018/Q3.csv
- /data/2018/Q4.csv
- /data/2019/Q1.csv

All files store data in the following format:
id,f1,f2,I
1,1,2,0
2,1,1,1
3,2,1,0
4,2,2,1
You run the following code:

```
data_store = Datastore.register_azure_blob_container(workspace=ws,
 datastore_name= 'data_store',
 container_name= 'quarterly_data',
 account_name='companydata',
 account_key='NRPxk8duxbM3…'
 create_if_not_exists=False)
```

You need to create a dataset named training_data and load the data from all files into a single data frame by using the following code:

```
data_frame = training_data.to_pandas_dataframe()
```

Solution: Run the following code:

```
from azureml.core import Dataset
paths = (data_store, 'data/*/*.csv')
training_data = Dataset.Tabular.from_delimited_files(paths)
```

Does the solution meet the goal?

- A. Yes
- B. No

Correct Answer: *A*
Define paths with two file paths instead.
Use Dataset.Tabular_from_delimeted as the data isn't cleansed.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-create-register-datasets

Question #48*Topic 3*

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 create an Azure Machine Learning service datastore in a workspace. The datastore contains the following files:
- /data/2018/Q1.csv
- /data/2018/Q2.csv
- /data/2018/Q3.csv
- /data/2018/Q4.csv
- /data/2019/Q1.csv

All files store data in the following format:
id,f1,f2,I
1,1,2,0
2,1,1,1
3,2,1,0
4,2,2,1
You run the following code:

```
data_store = Datastore.register_azure_blob_container(workspace=ws,
 datastore_name= 'data_store',
 container_name= 'quarterly_data',
 account_name='companydata',
 account_key='NRPxk8duxbM3…'
 create_if_not_exists=False)
```

You need to create a dataset named training_data and load the data from all files into a single data frame by using the following code:

```
data_frame = training_data.to_pandas_dataframe()
```

Solution: Run the following code:

```
from azureml.core import Dataset
paths = [(data_store, 'data/2018/*.csv'), (data_store, 'data/2019/*.csv')]
training_data = Dataset.File.from_files(paths)
```

Does the solution meet the goal?

- A. Yes
- B. No

Correct Answer: *B*
Use two file paths.
Use Dataset.Tabular_from_delimeted, instead of Dataset.File.from_files as the data isn't cleansed.
Note:
A FileDataset references single or multiple files in your datastores or public URLs. If your data is already cleansed, and ready to use in training experiments, you can download or mount the files to your compute as a FileDataset object.
A TabularDataset represents data in a tabular format by parsing the provided file or list of files. This provides you with the ability to materialize the data into a pandas or Spark DataFrame so you can work with familiar data preparation and training libraries without having to leave your notebook.
You can create a
TabularDataset object from .csv, .tsv, .parquet, .jsonl files, and from SQL query results.

Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-create-register-datasets

Question #49*Topic 3*

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 create an Azure Machine Learning service datastore in a workspace. The datastore contains the following files:
- /data/2018/Q1.csv
- /data/2018/Q2.csv
- /data/2018/Q3.csv
- /data/2018/Q4.csv
- /data/2019/Q1.csv

All files store data in the following format:
id,f1,f2,I
1,1,2,0
2,1,1,1
3,2,1,0
4,2,2,1
You run the following code:

```
data_store = Datastore.register_azure_blob_container(workspace=ws,
 datastore_name= 'data_store',
 container_name= 'quarterly_data',
 account_name='companydata',
 account_key='NRPxk8duxbM3...'
 create_if_not_exists=False)
```

You need to create a dataset named training_data and load the data from all files into a single data frame by using the following code:

```
data_frame = training_data.to_pandas_dataframe()
```

Solution: Run the following code:

```
from azureml.core import Dataset
paths = [(data_store, 'data/2018/*.csv'),(data_store, 'data/2019/*.csv')]
training_data = Dataset.Tabular.from_delimited_files(paths)
```

Does the solution meet the goal?

- A. Yes
- B. No

Correct Answer: *A*
Use two file paths.
Use Dataset.Tabular_from_delimeted as the data isn't cleansed.
Note:
A TabularDataset represents data in a tabular format by parsing the provided file or list of files. This provides you with the ability to materialize the data into a pandas or Spark DataFrame so you can work with familiar data preparation and training libraries without having to leave your notebook.

You can create a
TabularDataset object from .csv, .tsv, .parquet, .jsonl files, and from SQL query results.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-create-register-datasets

Question #50*Topic 3*

You plan to use the Hyperdrive feature of Azure Machine Learning to determine the optimal hyperparameter values when training a model.
You must use Hyperdrive to try combinations of the following hyperparameter values:
⇨ learning_rate: any value between 0.001 and 0.1
⇨ batch_size: 16, 32, or 64
You need to configure the search space for the Hyperdrive experiment.
Which two parameter expressions should you use? Each correct answer presents part of the solution.
NOTE: Each correct selection is worth one point.

- A. a choice expression for learning_rate
- B. a uniform expression for learning_rate
- C. a normal expression for batch_size
- D. a choice expression for batch_size
- E. a uniform expression for batch_size

Correct Answer: *BD*
B: Continuous hyperparameters are specified as a distribution over a continuous range of values.
Supported distributions include:
⇨ uniform(low, high) - Returns a value uniformly distributed between low and high
D: Discrete hyperparameters are specified as a choice among discrete values. choice can be: one or more comma-separated values

▪

⇨ a range object
⇨ any arbitrary list object
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-tune-hyperparameters

Question #51*Topic 3*

HOTSPOT -
Your Azure Machine Learning workspace has a dataset named real_estate_data. A sample of the data in the dataset follows.

postal_code	num_bedrooms	sq_feet	garage	price
12345	3	1300	0	23,9000
54321	1	950	0	11,0000
12346	2	1200	1	15,0000

You want to use automated machine learning to find the best regression model for predicting the price column.
You need to configure an automated machine learning experiment using the Azure Machine Learning SDK.
How should you complete the code? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.
Hot Area:

Answer Area

```
from azureml.core import Workspace
from azureml.core.compute import ComputeTarget
from azureml.core.runconfig import RunConfiguration
from azureml.train.automl import AutoMLConfig

ws = Workspace.from_config()
training_cluster = ComputeTarget(workspace=ws, name= 'aml-cluster1')
real_estate_ds = ws.datasets.get('real_estate_data')
split1_ds, split2_ds = real_estate_ds.random_split(percentage=0.7, seed=123)
automl_run_config = RunConfiguration(framework= "python")
automl_config = AutoMLConfig(        task= 'regression',
                                    compute_target= training_cluster,
                                    run_configuration=automl_run_config,
                                    primary_metric='r2_score',
                       [▼]          =split1_ds,
                       [▼]          =split2_ds
                       [▼]          ='price')
```

First drop-down (=split1_ds):

- X
- Y
- X_valid
- Y_valid
- training_data

Second drop-down (=split2_ds):

- X
- Y
- X_valid
- Y_valid
- validation_data

Third drop-down (='price')):

- y
- y_valid
- y_max
- label_column_name
- exclude_nan_labels

Correct Answer:

Answer Area

```
from azureml.core import Workspace
from azureml.core.compute import ComputeTarget
from azureml.core.runconfig import RunConfiguration
from azureml.train.automl import AutoMLConfig

ws = Workspace.from_config()
training_cluster = ComputeTarget(workspace=ws, name= 'aml-cluster1')
real_estate_ds = ws.datasets.get('real_estate_data')
split1_ds, split2_ds = real_estate_ds.random_split(percentage=0.7, seed=123)
automl_run_config = RunConfiguration(framework= "python")
automl_config = AutoMLConfig(          task= 'regression',
                                       compute_target= training_cluster,
                                       run_configuration=automl_run_config,
                                       primary_metric='r2_score',
                        [ ▼ ] =split1_ds,
                        [ ▼ ] =split2_ds
                        [ ▼ ] ='price')
```

Box 1 options (=split1_ds,):
- X
- Y
- X_valid
- Y_valid
- training_data

Box 2 options (=split2_ds):
- X
- Y
- X_valid
- Y_valid
- validation_data

Box 3 options (='price')):
- y
- y_valid
- y_max
- label_column_name
- exclude_nan_labels

Box 1: training_data -
The training data to be used within the experiment. It should contain both training features and a label column (optionally a sample weights column). If training_data is specified, then the label_column_name parameter must also be specified.

Box 2: validation_data -
Provide validation data: In this case, you can either start with a single data file and split it into training and validation sets or you can provide a separate data file for the validation set. Either way, the validation_data parameter in your AutoMLConfig object assigns which data to use as your validation set.
Example, the following code example explicitly defines which portion of the provided data in dataset to use for training and validation. dataset = Dataset.Tabular.from_delimited_files(data)
training_data, validation_data = dataset.random_split(percentage=0.8, seed=1) automl_config = AutoMLConfig(compute_target = aml_remote_compute, task = 'classification', primary_metric = 'AUC_weighted', training_data = training_data, validation_data = validation_data, label_column_name = 'Class'
)

Box 3: label_column_name -

label_column_name:
The name of the label column. If the input data is from a pandas.DataFrame which doesn't have column names, column indices can be used instead, expressed as integers.
This parameter is applicable to training_data and validation_data parameters.
Incorrect Answers:
X: The training features to use when fitting pipelines during an experiment. This setting is being deprecated. Please use training_data and label_column_name instead.
Y: The training labels to use when fitting pipelines during an experiment. This is the value your model will predict. This setting is being deprecated. Please use training_data and label_column_name instead.
X_valid: Validation features to use when fitting pipelines during an experiment.
If specified, then y_valid or sample_weight_valid must also be specified.
Y_valid: Validation labels to use when fitting pipelines during an experiment.
Both X_valid and y_valid must be specified together.
exclude_nan_labels: Whether to exclude rows with NaN values in the label. The default is True.
y_max: y_max (float)
Maximum value of y for a regression experiment. The combination of y_min and y_max are used to normalize test set metrics based on the input data range. If not specified, the maximum value is inferred from the data.
Reference:
https://docs.microsoft.com/en-us/python/api/azureml-train-automl-client/azureml.train.automl.automlconfig.automlconfig?view=azure-ml-py
Question #52*Topic 3*

HOTSPOT -
You have a multi-class image classification deep learning model that uses a set of labeled photographs. You create the following code to select hyperparameter values when training the model.

```
from azureml.train.hyperdrive import BayesianParameterSampling
param_sampling = BayesianParametersSampling ({
 "learning_rate": uniform(0.01, 0.1),
 "batch_size": choice(16, 32, 64, 128)}
)
```

For each of the following statements, select Yes if the statement is true. Otherwise, select No.
NOTE: Each correct selection is worth one point.
Hot Area:

Answer Area

	Yes	No
Hyperparameter combinations for the runs are selected based on how previous samples performed in the previous experiment run.	○	○
The learning rate value 0.09 might be used during model training.	○	○
You can define an early termination policy for this hyperparameter tuning run.	○	○

Correct Answer:

Answer Area

	Yes	No
Hyperparameter combinations for the runs are selected based on how previous samples performed in the previous experiment run.	○	○
The learning rate value 0.09 might be used during model training.	○	○
You can define an early termination policy for this hyperparameter tuning run.	○	○

Box 1: Yes -
Hyperparameters are adjustable parameters you choose to train a model that govern the training process itself. Azure Machine Learning allows you to automate hyperparameter exploration in an efficient manner, saving you significant time and resources. You specify the range of hyperparameter values and a maximum number of training runs. The system then automatically launches multiple simultaneous runs with different parameter configurations and finds the configuration that results in the best performance, measured by the metric you choose. Poorly performing training runs are automatically early terminated, reducing wastage of compute resources. These resources are instead used to explore other hyperparameter configurations.

Box 2: Yes -
uniform(low, high) - Returns a value uniformly distributed between low and high

Box 3: No -
Bayesian sampling does not currently support any early termination policy.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-tune-hyperparameters
Question #53*Topic 3*

You run an automated machine learning experiment in an Azure Machine Learning workspace. Information about the run is listed in the table below:

Experiment	Run ID	Status	Created on	Duration
auto_ml_clasification	AutoML_1234567890-123	Completed	11/11/2019 11:00:00 AM	00:27:11

You need to write a script that uses the Azure Machine Learning SDK to retrieve the best iteration of the experiment run.
Which Python code segment should you use?
A.

```
from azureml.core import Workspace
from azureml.train.automl.run import AutoMLRun
ws = Workspace.from_config()
automl_ex = ws.experiments.get('auto_ml_classification')
best_iter = automl_ex.archived_time.find('11/11/2019 11:00:00 AM')
```

B.

```
from azureml.core import Workspace
from azureml.train.automl.run import AutoMLRun
automl_ex = ws.experiments.get('auto_ml_classification')
automl_run = AutoMLRun(automl_ex, 'AutoML_1234567890-123')
best_iter = automl_run.current_run
```

C.

```
from azureml.core import Workspace
from azureml.train.automl.run import AutoMLRun
ws = Workspace.from_config()
automl_ex = ws.experiments.get('auto_ml_classification')
best_iter = list(automl_ex.get_runs())[0]
```

D.

```
from azureml.core import Workspace
from azureml.train.automl.run import AutoMLRun
ws = Workspace.from_config()
automl_ex = ws.experiments.get('auto_ml_classification')
automl_run = AutoMLRun(automl_ex, 'AutoML_1234567890-123')
best_iter = automl_run.get_output()[0]
```

E.

```
from azureml.core import Workspace
from azureml.train.automl.run import AutoMLRun
ws = Workspace.from_config()
automl_ex = ws.experiments.get('auto_ml_classification')
best_iter = automl_ex.get_runs('AutoML_1234567890-123')
```

Correct Answer: *D*
The get_output method on automl_classifier returns the best run and the fitted model for the last invocation. Overloads on get_output allow you to retrieve the best run and fitted model for any logged metric or for a particular iteration.
In [ ]:
best_run, fitted_model = local_run.get_output()
Reference:
https://notebooks.azure.com/azureml/projects/azureml-getting-started/html/how-to-use-azureml/automated-machine-learning/classification-with-deployment/auto- ml-classification-with-deployment.ipynb
Question #54*Topic 3*

You have a comma-separated values (CSV) file containing data from which you want to train a classification model.
You are using the Automated Machine Learning interface in Azure Machine Learning studio to train the classification model. You set the task type to Classification.
You need to ensure that the Automated Machine Learning process evaluates only linear models.
What should you do?

- A. Add all algorithms other than linear ones to the blocked algorithms list.
- B. Set the Exit criterion option to a metric score threshold.
- C. Clear the option to perform automatic featurization.
- D. Clear the option to enable deep learning.
- E. Set the task type to Regression.

Correct Answer: *A*
Automatic featurization can fit non-linear models.
Reference:
https://econml.azurewebsites.net/spec/estimation/dml.html
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-use-automated-ml-for-ml-models

Question #55*Topic 3*

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 plan to use a Python script to run an Azure Machine Learning experiment. The script creates a reference to the experiment run context, loads data from a file, identifies the set of unique values for the label column, and completes the experiment run:

```
from azureml.core import Run
import pandas as pd

run = Run.get_context()
data = pd.read_csv('data.csv')
label_vals = data['label'].unique()
# Add code to record metrics here
run.complete()
```

The experiment must record the unique labels in the data as metrics for the run that can be reviewed later.
You must add code to the script to record the unique label values as run metrics at the point indicated by the comment.
Solution: Replace the comment with the following code:
run.upload_file('outputs/labels.csv', './data.csv')
Does the solution meet the goal?

- A. Yes
- B. No

Correct Answer: *B*
label_vals has the unique labels (from the statement label_vals = data['label'].unique()), and it has to be logged.
Note:
Instead use the run_log function to log the contents in label_vals: for label_val in label_vals:
run.log('Label Values', label_val)
Reference:
https://www.element61.be/en/resource/azure-machine-learning-services-complete-toolbox-ai

Question #56*Topic 3*

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 plan to use a Python script to run an Azure Machine Learning experiment. The script creates a reference to the experiment run context, loads data from a file, identifies the set of unique values for the label column, and completes the experiment run:

```
from azureml.core import Run
import pandas as pd

run = Run.get_context()
data = pd.read_csv('data.csv')
label_vals = data['label'].unique()
# Add code to record metrics here
run.complete()
```

The experiment must record the unique labels in the data as metrics for the run that can be reviewed later.
You must add code to the script to record the unique label values as run metrics at the point indicated by the comment.
Solution: Replace the comment with the following code:
run.log_table('Label Values', label_vals)
Does the solution meet the goal?

- A. Yes
- B. No

Correct Answer: *B*
Instead use the run_log function to log the contents in label_vals: for label_val in label_vals:
run.log('Label Values', label_val)
Reference:
https://www.element61.be/en/resource/azure-machine-learning-services-complete-toolbox-ai
Question #57*Topic 3*

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 plan to use a Python script to run an Azure Machine Learning experiment. The script creates a reference to the experiment run context, loads data from a file, identifies the set of unique values for the label column, and completes the experiment run:

```
from azureml.core import Run
import pandas as pd

run = Run.get_context()
data = pd.read_csv('data.csv')
label_vals = data['label'].unique()
# Add code to record metrics here
run.complete()
```

The experiment must record the unique labels in the data as metrics for the run that can be reviewed later.
You must add code to the script to record the unique label values as run metrics at the point indicated by the comment.
Solution: Replace the comment with the following code:
for label_val in label_vals:
run.log('Label Values', label_val)
Does the solution meet the goal?

- A. Yes
- B. No

Correct Answer: *B*
The run_log function is used to log the contents in label_vals: for label_val in label_vals:
run.log('Label Values', label_val)
Reference:
https://www.element61.be/en/resource/azure-machine-learning-services-complete-toolbox-ai

Question #58*Topic 3*

HOTSPOT -
You publish a batch inferencing pipeline that will be used by a business application.
The application developers need to know which information should be submitted to and returned by the REST interface for the published pipeline.
You need to identify the information required in the REST request and returned as a response from the published pipeline.
Which values should you use in the REST request and to expect in the response? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

REST Request	Value
Request Header	JSON containing the run ID JSON containing the pipeline ID JSON containing the experiment name JSON containing an OAuth bearer token
Request Body	JSON containing the run ID JSON containing the pipeline ID JSON containing the experiment name JSON containing an OAuth bearer token
Response	JSON containing the run ID JSON containing a list of predictions JSON containing the experiment name JSON containing a path to the parallel_run_step.txt output file

Correct Answer:

Answer Area

REST Request	Value
Request Header	JSON containing the run ID JSON containing the pipeline ID JSON containing the experiment name JSON containing an OAuth bearer token
Request Body	JSON containing the run ID JSON containing the pipeline ID JSON containing the experiment name JSON containing an OAuth bearer token
Response	JSON containing the run ID JSON containing a list of predictions JSON containing the experiment name JSON containing a path to the parallel_run_step.txt output file

Box 1: JSON containing an OAuth bearer token
Specify your authentication header in the request.
To run the pipeline from the REST endpoint, you need an OAuth2 Bearer-type authentication header.
Box 2: JSON containing the experiment name
Add a JSON payload object that has the experiment name.
Example:

```
rest_endpoint = published_pipeline.endpoint
response = requests.post(rest_endpoint,
headers=auth_header,
json={"ExperimentName": "batch_scoring",
"ParameterAssignments": {"process_count_per_node": 6}})
run_id = response.json()["Id"]
```

Box 3: JSON containing the run ID
Make the request to trigger the run. Include code to access the Id key from the response dictionary to get the value of the run ID.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/tutorial-pipeline-batch-scoring-classification
Question #59*Topic 3*

HOTSPOT -

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

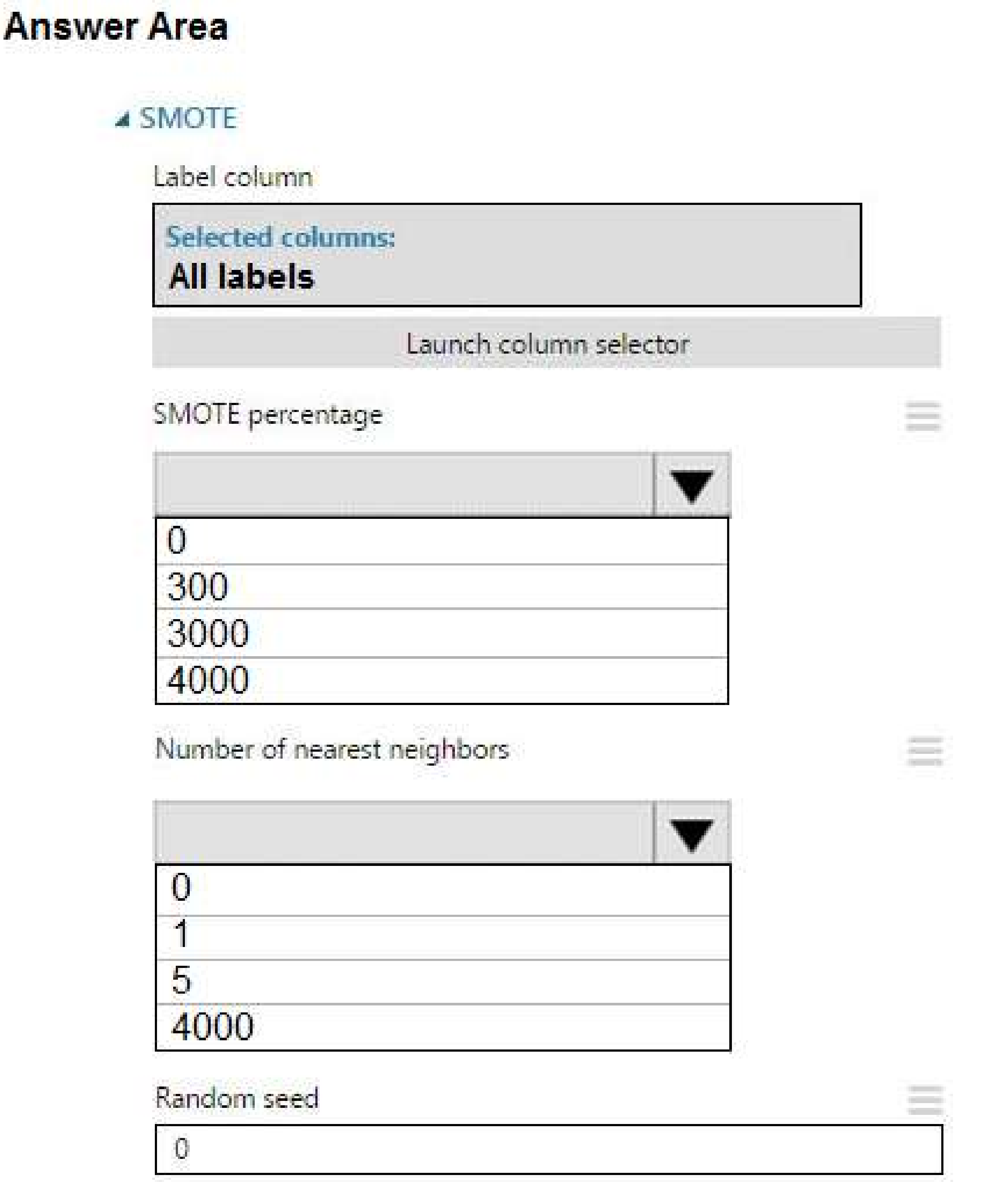
Synthetic Minority Oversampling Technique (SMOTE) module to the experiment.

You need to configure the module.

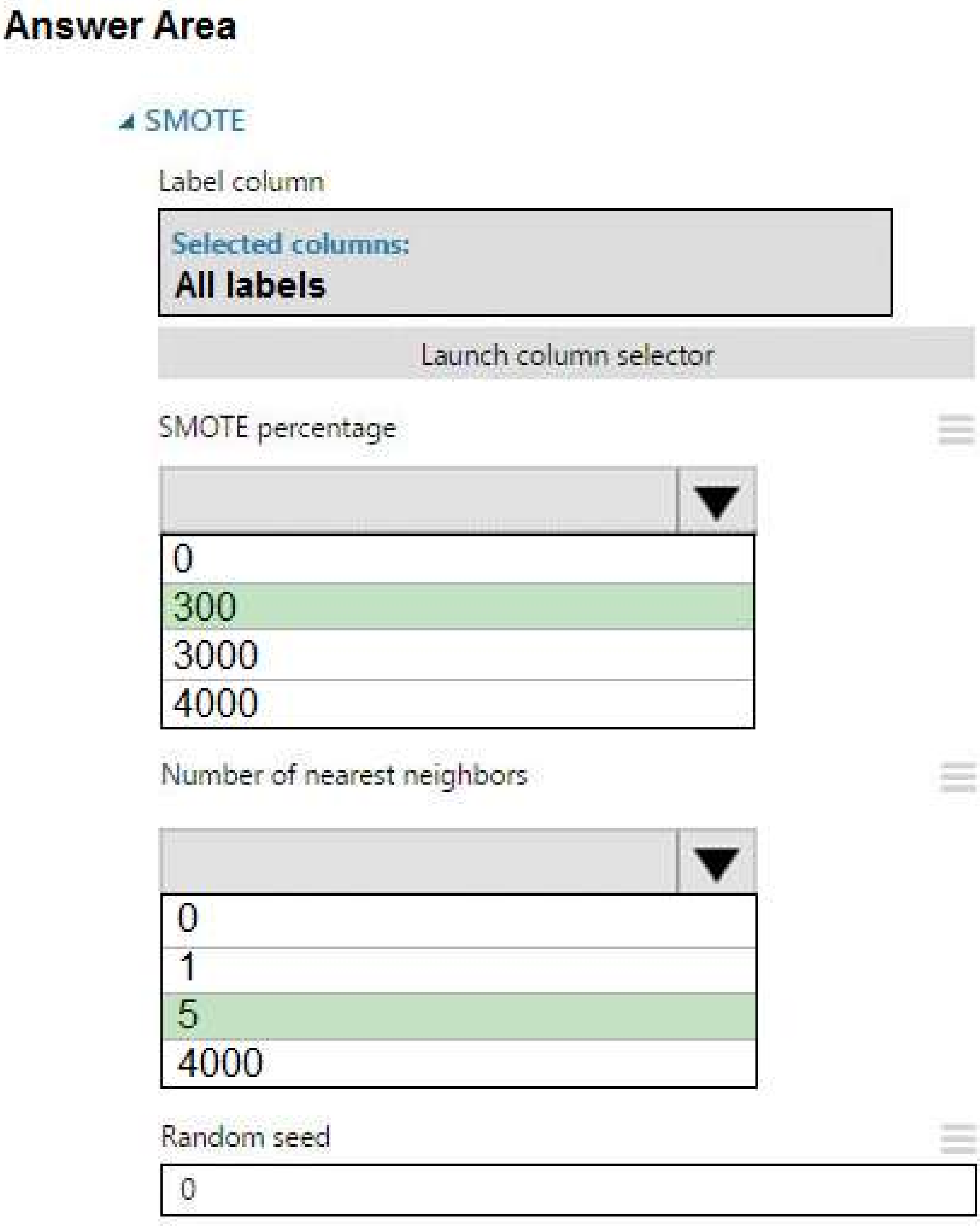
Which values should you use? To answer, select the appropriate options in the dialog box in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:



Correct Answer:



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.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/smote
Question #60*Topic 3*

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=0.01
- C. k=5
- D. k=1

Correct Answer: *B*
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.

Question #61*Topic 3*
HOTSPOT -
You are running Python code interactively in a Conda environment. The environment includes all required Azure Machine Learning SDK and MLflow packages.
You must use MLflow to log metrics in an Azure Machine Learning experiment named mlflow-experiment.
How should you complete the code? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

```
import mlflow
from azureml.core import Workspace
ws = Workspace.from_config()
# Set the MLflow logging target
```

mlflow.tracking.client = ws
mlflow.set_tracking_uri(ws.get_mlflow_tracking_uri())
mlflow.log_param('workspace', ws)

```
# Configure experiment
```

mlflow-experiment = Run.get_context()
mlflow.get_run('mlflow-experiment')
mlflow.set_experiment('mlflow-experiment')

```
# Begin the experiment run
with
```

mlflow.active_run
mlflow.start_run()
Run.get_context()

```
    # Log my_metric with value 1.00
    ('my_metric', 1.00)
```

run.log()
mlflow.log_metric
print

```
print("Finished!")
```

Correct Answer:

Answer Area

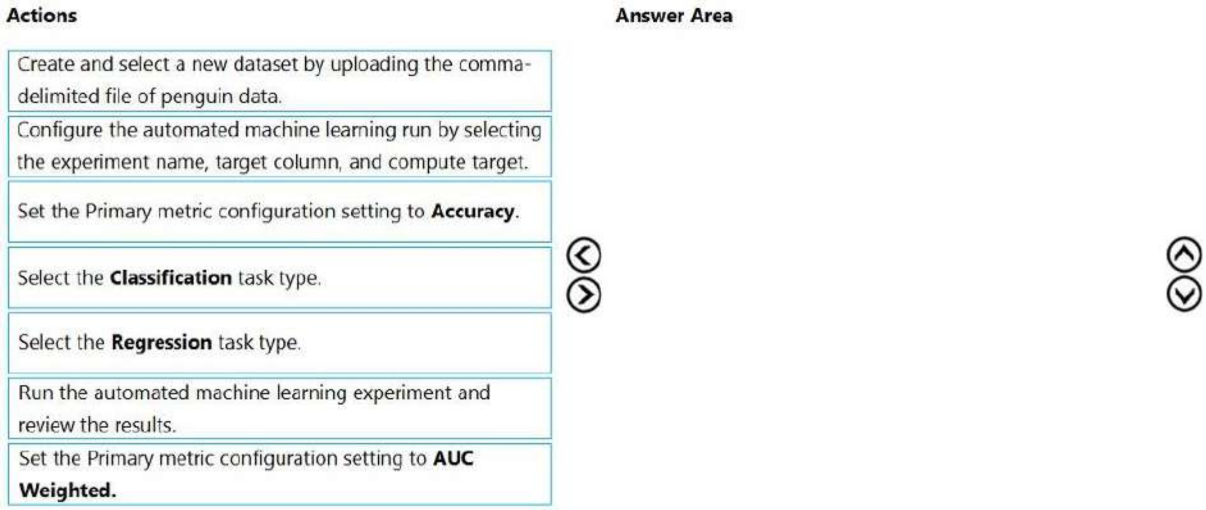
```
import mlflow
from azureml.core import Workspace
ws = Workspace.from_config()
# Set the MLflow logging target
[ ▼ ]
mlflow.tracking.client = ws
mlflow.set_tracking_uri(ws.get_mlflow_tracking_uri())
mlflow.log_param('workspace', ws)
# Configure experiment
[ ▼ ]
mlflow-experiment = Run.get_context()
mlflow.get_run('mlflow-experiment')
mlflow.set_experiment('mlflow-experiment')
# Begin the experiment run
with [ ▼ ]
     mlflow.active_run
     mlflow.start_run()
     Run.get_context()
    # Log my_metric with value 1.00
    [ ▼ ] ('my_metric', 1.00)
    run.log()
    mlflow.log_metric
    print
print("Finished!")
```

Box 1: mlflow.set_tracking_uri(ws.get_mlflow_tracking_uri())
In the following code, the get_mlflow_tracking_uri() method assigns a unique tracking URI address to the workspace, ws, and set_tracking_uri() points the MLflow tracking URI to that address.
mlflow.set_tracking_uri(ws.get_mlflow_tracking_uri())
Box 2: mlflow.set_experiment(experiment_name)
Set the MLflow experiment name with set_experiment() and start your training run with start_run().
Box 3: mlflow.start_run()

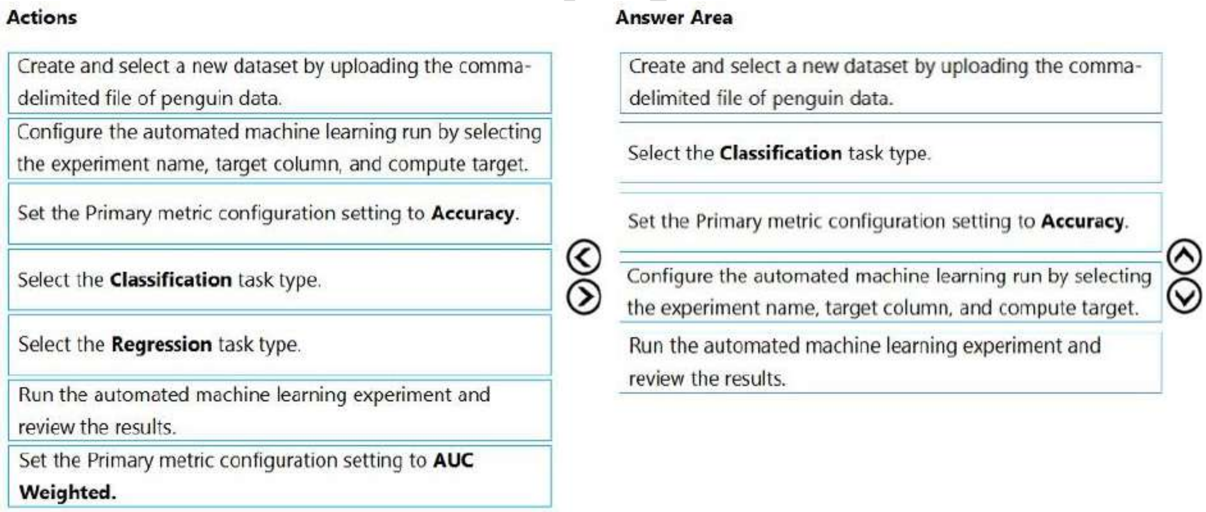
Box 4: mlflow.log_metric -
Then use log_metric() to activate the MLflow logging API and begin logging your training run metrics.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-use-mlflow
Question #62*Topic 3*

DRAG DROP -
You are creating a machine learning model that can predict the species of a penguin from its measurements. You have a file that contains measurements for three species of penguin in comma-delimited format.

The model must be optimized for area under the received operating characteristic curve performance metric, averaged for each class.
You need to use the Automated Machine Learning user interface in Azure Machine Learning studio to run an experiment and find the best performing model.
Which five 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.
Select and Place:



Correct Answer:



Step 1:Create and select a new dataset by uploading he command-delimited file of penguin data.
Step 2: Select the Classification task type
Step 3: Set the Primary metric configuration setting to Accuracy.
The available metrics you can select is determined by the task type you choose.
Primary metrics for classification scenarios:
Post thresholded metrics, like accuracy, average_precision_score_weighted, norm_macro_recall, and precision_score_weighted may not optimize as well for datasets which are very small, have very large class skew (class imbalance), or when the expected metric value is very close to 0.0 or 1.0. In those cases,
AUC_weighted can be a better choice for the primary metric.
Step 4: Configure the automated machine learning run by selecting the experiment name, target

column, and compute target
Step 5: Run the automated machine learning experiment and review the results.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-configure-auto-train
Question #63*Topic 3*

HOTSPOT -
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.
Hot Area:

Answer Area

Question	Answer Choise
Which H value should you select based on the data?	1 2 3 4 5
What H value displays the poorest training result?	1 2 3 4 5

Correct Answer:

Answer Area

Question	Answer Choise
Which H value should you select based on the data?	1 2 3 4 5
What H value displays the poorest training result?	1 2 3 4 5

Box 1: 4 -
Choose the one which has lower training and validation error and also the closest match.
Minimize variance (difference between validation error and train error).

Box 2: 5 -
Minimize variance (difference between validation error and train error).
Reference:
https://medium.com/comet-ml/organizing-machine-learning-projects-project-management-guidelines-2d2b85651bbd
Question #64*Topic 3*

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 create a model to forecast weather conditions based on historical data.
You need to create a pipeline that runs a processing script to load data from a datastore and pass the processed data to a machine learning model training script.
Solution: Run the following code:

```
datastore = ws.get_default_datastore()
data_output = pd.read_csv("traindata.csv")
process_step = PythonScriptStep(script_name="process.py",
 arguments=["--data_for_train", data_output],
 outputs=[data_output],compute_target=aml_compute,
 source_directory=process_directory)
train_step = PythonScriptStep(script_name="train.py",
 arguments=["--data_for_train", data_output],
 inputs=[data_output],compute_target=aml_compute,
 source_directory=train_directory)
pipeline = Pipeline(workspace=ws, steps=[process_step, train_step])
```

Does the solution meet the goal?

- A. Yes
- B. No

Correct Answer: *B*
The two steps are present: process_step and train_step
The training data input is not setup correctly.
Note:
Data used in pipeline can be produced by one step and consumed in another step by providing a PipelineData object as an output of one step and an input of one or more subsequent steps.
PipelineData objects are also used when constructing Pipelines to describe step dependencies. To specify that a step requires the output of another step as input, use a PipelineData object in the constructor of both steps.
For example, the pipeline train step depends on the process_step_output output of the pipeline process step: from azureml.pipeline.core import Pipeline, PipelineData from azureml.pipeline.steps import PythonScriptStep datastore = ws.get_default_datastore() process_step_output = PipelineData("processed_data", datastore=datastore) process_step = PythonScriptStep(script_name="process.py", arguments=["--data_for_train", process_step_output], outputs=[process_step_output], compute_target=aml_compute, source_directory=process_directory) train_step = PythonScriptStep(script_name="train.py", arguments=["--data_for_train", process_step_output], inputs=[process_step_output], compute_target=aml_compute, source_directory=train_directory) pipeline = Pipeline(workspace=ws, steps=[process_step, train_step])
Reference:
https://docs.microsoft.com/en-us/python/api/azureml-pipeline-core/azureml.pipeline.core.pipelinedata?view=azure-ml-py
Question #65*Topic 3*

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 create a model to forecast weather conditions based on historical data.
You need to create a pipeline that runs a processing script to load data from a datastore and pass the processed data to a machine learning model training script.
Solution: Run the following code:

```
datastore = ws.get_default_datastore()
data_output = PipelineData("processed_data", datastore=datastore)
process_step = PythonScriptStep(script_name="process.py",
 arguments=["--data_for_train", data_output],
 outputs=[data_output], compute_target=aml_compute,
 source_directory=process_directory)
pipeline = Pipeline(workspace=ws, steps=[process_step])
```

Does the solution meet the goal?

- A. Yes
- B. No

Correct Answer: *B*
train_step is missing.
Reference:
https://docs.microsoft.com/en-us/python/api/azureml-pipeline-core/azureml.pipeline.core.pipelinedata?view=azure-ml-py

Question #66*Topic 3*

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 create a model to forecast weather conditions based on historical data.
You need to create a pipeline that runs a processing script to load data from a datastore and pass the processed data to a machine learning model training script.
Solution: Run the following code:

```
datastore = ws.get_default_datastore()
data_input = PipelineData("raw_data", datastore=rawdatastore)
data_output = PipelineData("processed_data", datastore=datastore)
process_step = PythonScriptStep(script_name="process.py",
 arguments=["--data_for_train", data_input],
 outputs=[data_output], compute_target=aml_compute,
 source_directory=process_directory)
train_step = PythonScriptStep(script_name="train.py",
 arguments=["--data_for_train", data_input], inputs=[data_output],
 compute_target=aml_compute, source_directory=train_directory)
pipeline = Pipeline(workspace=ws, steps=[process_step, train_step])
```

Does the solution meet the goal?

- A. Yes
- B. No

Correct Answer: *B*
Note: Data used in pipeline can be produced by one step and consumed in another step by providing a PipelineData object as an output of one step and an input of one or more subsequent steps.
Compare with this example, the pipeline train step depends on the process_step_output output of the pipeline process step: from azureml.pipeline.core import Pipeline, PipelineData from azureml.pipeline.steps import PythonScriptStep datastore = ws.get_default_datastore() process_step_output = PipelineData("processed_data", datastore=datastore) process_step = PythonScriptStep(script_name="process.py", arguments=["--data_for_train", process_step_output], outputs=[process_step_output], compute_target=aml_compute, source_directory=process_directory) train_step = PythonScriptStep(script_name="train.py", arguments=["--data_for_train", process_step_output], inputs=[process_step_output], compute_target=aml_compute, source_directory=train_directory) pipeline = Pipeline(workspace=ws, steps=[process_step, train_step])
Reference:
https://docs.microsoft.com/en-us/python/api/azureml-pipeline-core/azureml.pipeline.core.pipelinedata?view=azure-ml-py
Question #67*Topic 3*

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 have a Python script named train.py in a local folder named scripts. The script trains a regression model by using scikit-learn. The script includes code to load a training data file which is also located in the scripts folder.
You must run the script as an Azure ML experiment on a compute cluster named aml-compute.
You need to configure the run to ensure that the environment includes the required packages for model training. You have instantiated a variable named aml- compute that references the target compute cluster.
Solution: Run the following code:

```
from azureml.train.sklearn import SKLearn
sk_est = SKLearn(source_directory='./scripts',
 compute_target=aml-compute,
 entry_script='train.py')
```

Does the solution meet the goal?

- A. Yes
- B. No

Correct Answer: B
The scikit-learn estimator provides a simple way of launching a scikit-learn training job on a compute target. It is implemented through the SKLearn class, which can be used to support single-node CPU training.
Example:
from azureml.train.sklearn import SKLearn
}
estimator = SKLearn(source_directory=project_folder,

compute_target=compute_target,
entry_script='train_iris.py'
)
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-train-scikit-learn

Question #68*Topic 3*

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 have a Python script named train.py in a local folder named scripts. The script trains a regression model by using scikit-learn. The script includes code to load a training data file which is also located in the scripts folder.
You must run the script as an Azure ML experiment on a compute cluster named aml-compute.
You need to configure the run to ensure that the environment includes the required packages for model training. You have instantiated a variable named aml- compute that references the target compute cluster.
Solution: Run the following code:

```
from azureml.train.dnn import TensorFlow
sk_est = TensorFlow(source_directory='./scripts',
  compute_target=aml-compute,
  entry_script='train.py')
```

Does the solution meet the goal?

- A. Yes
- B. No

Correct Answer: *B*
The scikit-learn estimator provides a simple way of launching a scikit-learn training job on a compute target. It is implemented through the SKLearn class, which can be used to support single-node CPU training.
Example:
from azureml.train.sklearn import SKLearn
}
estimator = SKLearn(source_directory=project_folder,
compute_target=compute_target,
entry_script='train_iris.py'
)
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-train-scikit-learn
Question #69*Topic 3*

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 have a Python script named train.py in a local folder named scripts. The script trains a regression model by using scikit-learn. The script includes code to load a training data file which is also located in the scripts folder.
You must run the script as an Azure ML experiment on a compute cluster named aml-compute.
You need to configure the run to ensure that the environment includes the required packages for model training. You have instantiated a variable named aml- compute that references the target compute cluster.
Solution: Run the following code:

```
from azureml.train.estimator import Estimator
sk_est = Estimator(source_directory='./scripts',
 compute_target=aml-compute,
 entry_script='train.py',
 conda_packages=['scikit-learn'])
```

Does the solution meet the goal?

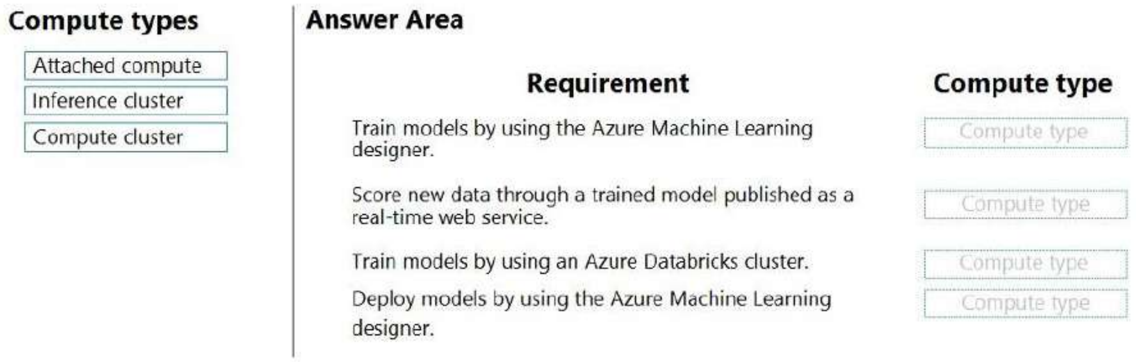
- A. Yes
- B. No

Correct Answer: *A*
The scikit-learn estimator provides a simple way of launching a scikit-learn training job on a compute target. It is implemented through the SKLearn class, which can be used to support single-node CPU training.
Example:
from azureml.train.sklearn import SKLearn
}
estimator = SKLearn(source_directory=project_folder,
compute_target=compute_target,
entry_script='train_iris.py'
)
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-train-scikit-learn

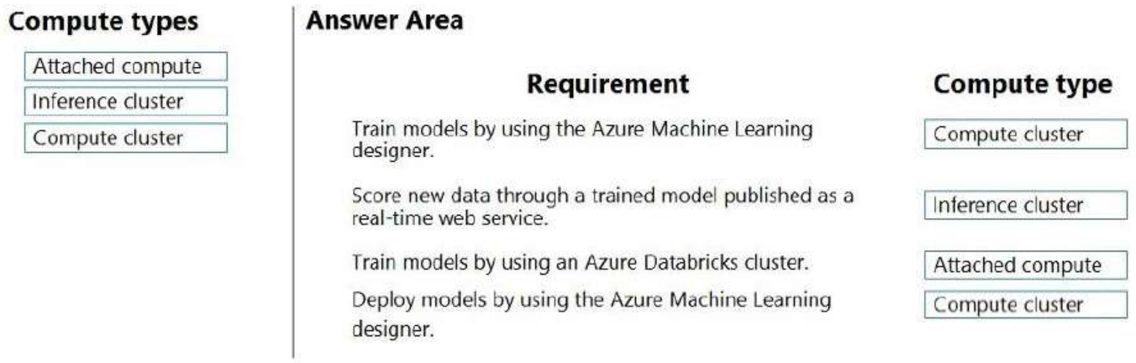
Question #70*Topic 3*

DRAG DROP -
You create machine learning models by using Azure Machine Learning.
You plan to train and score models by using a variety of compute contexts. You also plan to create a new compute resource in Azure Machine Learning studio.
You need to select the appropriate compute types.
Which compute types should you select? To answer, drag the appropriate compute types to the correct requirements. Each compute type 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.

Select and Place:



Correct Answer:



Box 1: Compute cluster -
Create a single or multi node compute cluster for your training, batch inferencing or reinforcement learning workloads.

Box 2: Inference cluster -

Box 3: Attached compute -
The compute types that can currently be attached for training include:

A remote VM -
Azure Databricks (for use in machine learning pipelines)
Azure Data Lake Analytics (for use in machine learning pipelines)

Azure HDInsight -

Box 4: Compute cluster -
Note: There are four compute types:

Compute instance -

Compute clusters -

Inference clusters -

Attached compute -

Note 2:

Compute clusters -
Create a single or multi node compute cluster for your training, batch inferencing or reinforcement learning workloads.

Attached compute -
To use compute targets created outside the Azure Machine Learning workspace, you must attach them. Attaching a compute target makes it available to your workspace. Use Attached compute to attach a compute target for training. Use Inference clusters to attach an AKS cluster for inferencing.

Inference clusters -
Create or attach an Azure Kubernetes Service (AKS) cluster for large scale inferencing.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-create-attach-compute-studio
Question #71*Topic 3*

DRAG DROP -
You are building an experiment using the Azure Machine Learning designer.
You split a dataset into training and testing sets. You select the Two-Class Boosted Decision Tree as the algorithm.
You need to determine the Area Under the Curve (AUC) of the model.
Which three 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.
Select and Place:

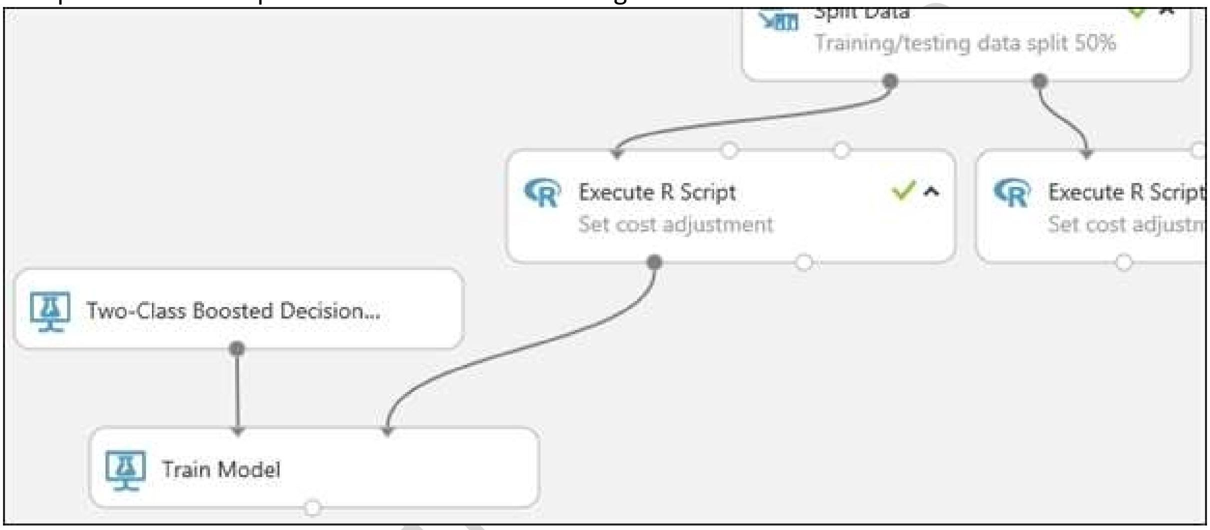
Modules	Answer Area
Export Data	
Tune Model Hyperparameters	
Cross Validate Model	
Evaluate Model	
Score Model	
Train Model	

Correct Answer:

Modules	Answer Area
Export Data	Train Model
Tune Model Hyperparameters	Score Model
Cross Validate Model	Evaluate Model
Evaluate Model	
Score Model	
Train Model	

Step 1: Train Model -

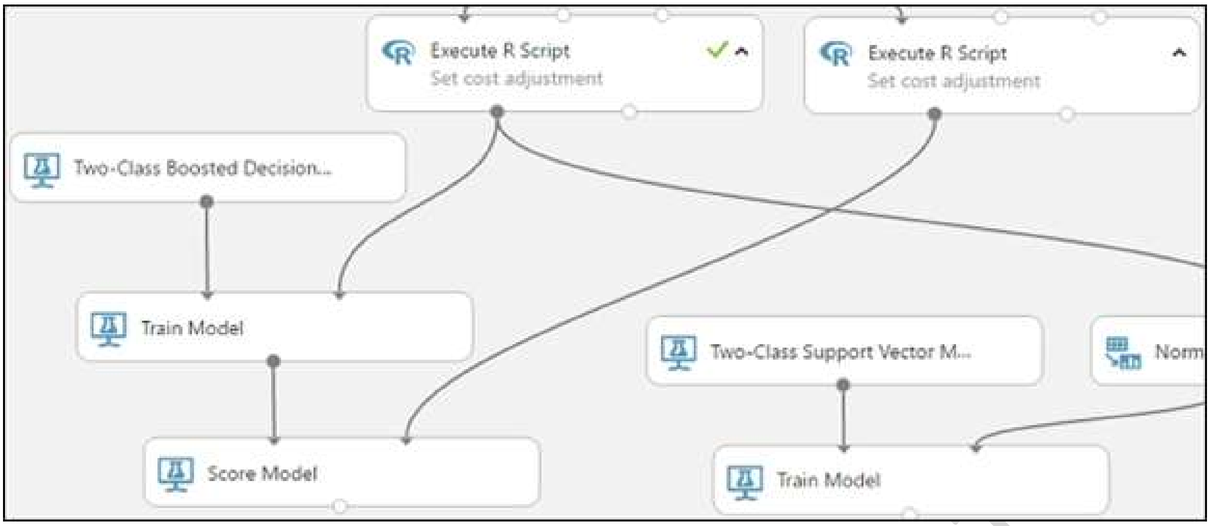
Two-Class Boosted Decision Tree -
First, set up the boosted decision tree model.
1. Find the Two-Class Boosted Decision Tree module in the module palette and drag it onto the canvas.
2. Find the Train Model module, drag it onto the canvas, and then connect the output of the Two-Class Boosted Decision Tree module to the left input port of the
Train Model module.
The Two-Class Boosted Decision Tree module initializes the generic model, and Train Model uses training data to train the model.
3. Connect the left output of the left Execute R Script module to the right input port of the Train Model module (in this tutorial you used the data coming from the left side of the Split Data module for training).
This portion of the experiment now looks something like this:



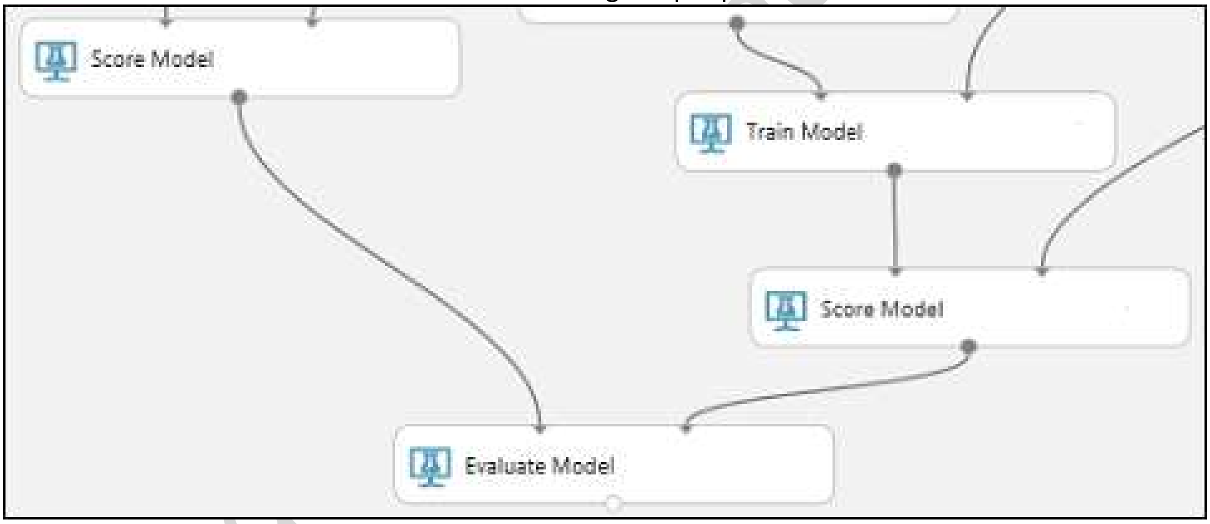
Step 2: Score Model -

Score and evaluate the models -
You use the testing data that was separated out by the Split Data module to score our trained models. You can then compare the results of the two models to see which generated better results.

Add the Score Model modules -
1. Find the Score Model module and drag it onto the canvas.
2. Connect the Train Model module that's connected to the Two-Class Boosted Decision Tree module to the left input port of the Score Model module.
3. Connect the right Execute R Script module (our testing data) to the right input port of the Score Model module.



Step 3: Evaluate Model -
To evaluate the two scoring results and compare them, you use an Evaluate Model module.
1. Find the Evaluate Model module and drag it onto the canvas.
2. Connect the output port of the Score Model module associated with the boosted decision tree model to the left input port of the Evaluate Model module.
3. Connect the other Score Model module to the right input port.



Question #72*Topic 3*

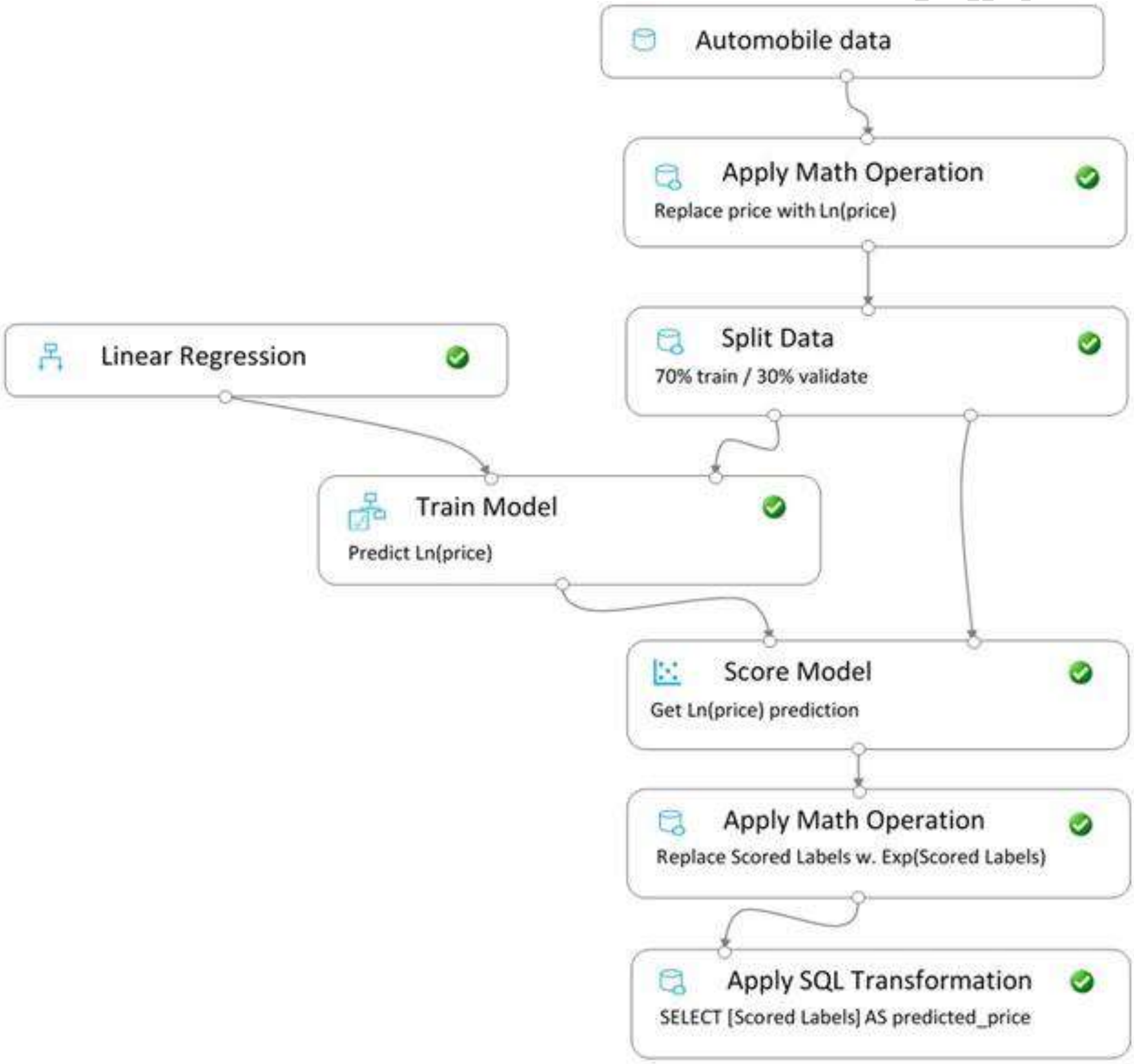
You create a multi-class image classification deep learning model that uses a set of labeled images.
You create a script file named train.py that uses the PyTorch
1.3 framework to train the model.
You must run the script by using an estimator. The code must not require any additional Python libraries to be installed in the environment for the estimator. The time required for model training must be minimized.
You need to define the estimator that will be used to run the script.
Which estimator type should you use?

- A. TensorFlow
- B. PyTorch
- C. SKLearn
- D. Estimator

Correct Answer: *B*
For PyTorch, TensorFlow and Chainer tasks, Azure Machine Learning provides respective PyTorch, TensorFlow, and Chainer estimators to simplify using these frameworks.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-train-ml-models
Question #73*Topic 3*

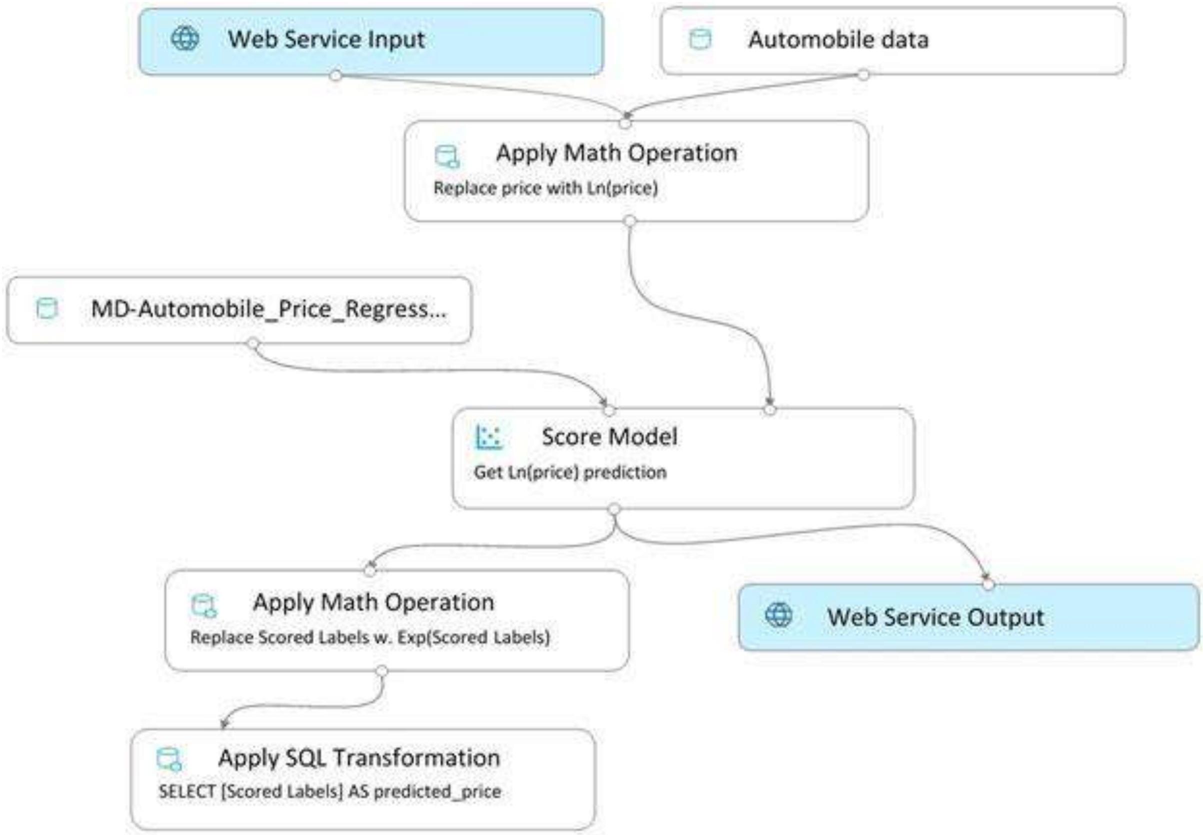
You create a pipeline in designer to train a model that predicts automobile prices.
Because of non-linear relationships in the data, the pipeline calculates the natural log (Ln) of the prices in the training data, trains a model to predict this natural log of price value, and then calculates the exponential of the scored label to get the predicted price.
The training pipeline is shown in the exhibit. (Click the Training pipeline tab.)

Training pipeline -



You create a real-time inference pipeline from the training pipeline, as shown in the exhibit. (Click the Real-time pipeline tab.)

Real-time pipeline -



You need to modify the inference pipeline to ensure that the web service returns the exponential of the scored label as the predicted automobile price and that client applications are not required to include a price value in the input values.
Which three modifications must you make to the inference pipeline? Each correct answer presents part of the solution.
NOTE: Each correct selection is worth one point.

- A. Connect the output of the Apply SQL Transformation to the Web Service Output module.
- B. Replace the Web Service Input module with a data input that does not include the price column.
- C. Add a Select Columns module before the Score Model module to select all columns other than price.
- D. Replace the training dataset module with a data input that does not include the price column.
- E. Remove the Apply Math Operation module that replaces price with its natural log from the data flow.
- F. Remove the Apply SQL Transformation module from the data flow.

Correct Answer: *ACE*

Question #74*Topic 3*

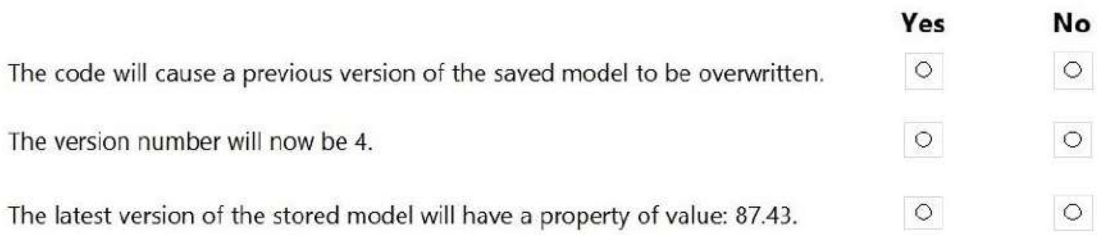
HOTSPOT -
You register the following versions of a model.

Model name	Model version	Tags	Properties
healthcare_model	3	'Training context':'CPU Compute'	value:87.43
healthcare_model	2	'Training context':'CPU Compute'	value:54.98
healthcare_model	1	'Training context':'CPU Compute'	value:23.56

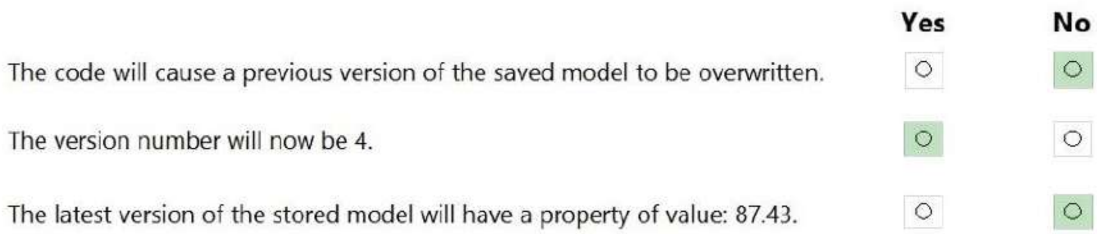
You use the Azure ML Python SDK to run a training experiment. You use a variable named run to reference the experiment run.
After the run has been submitted and completed, you run the following code:

```
run.register_model(model_path='outputs/model.pkl',
 model_name='healthcare_model',
 tags={'Training context':'CPU Compute'} )
```

For each of the following statements, select Yes if the statement is true. Otherwise, select No.
NOTE: Each correct selection is worth one point.
Hot Area:
Answer Area



Correct Answer:
Answer Area



Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-deploy-and-where
Question #75*Topic 3*

You are creating a classification model for a banking company to identify possible instances of credit card fraud. You plan to create the model in Azure Machine Learning by using automated machine learning.
The training dataset that you are using is highly unbalanced.
You need to evaluate the classification model.
Which primary metric should you use?

- A. normalized_mean_absolute_error
- B. AUC_weighted
- C. accuracy
- D. normalized_root_mean_squared_error
- E. spearman_correlation

Correct Answer: *B*
AUC_weighted is a Classification metric.
Note: AUC is the Area under the Receiver Operating Characteristic Curve. Weighted is the arithmetic mean of the score for each class, weighted by the number of true instances in each class.
Incorrect Answers:
A: normalized_mean_absolute_error is a regression metric, not a classification metric.
C: When comparing approaches to imbalanced classification problems, consider using metrics beyond accuracy such as recall, precision, and AUROC. It may be that switching the metric you optimize for during parameter selection or model selection is enough to provide desirable performance detecting the minority class.
D: normalized_root_mean_squared_error is a regression metric, not a classification metric.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-understand-automated-ml
Question #76*Topic 3*

You create a machine learning model by using the Azure Machine Learning designer. You publish the model as a real-time service on an Azure Kubernetes
Service (AKS) inference compute cluster. You make no changes to the deployed endpoint configuration.
You need to provide application developers with the information they need to consume the endpoint.
Which two values should you provide to application developers? Each correct answer presents part of the solution.
NOTE: Each correct selection is worth one point.

- A. The name of the AKS cluster where the endpoint is hosted.
- B. The name of the inference pipeline for the endpoint.
- C. The URL of the endpoint.
- D. The run ID of the inference pipeline experiment for the endpoint.
- E. The key for the endpoint.

Correct Answer: *CE*
Deploying an Azure Machine Learning model as a web service creates a REST API endpoint. You can send data to this endpoint and receive the prediction returned by the model.
You create a web service when you deploy a model to your local environment, Azure Container Instances, Azure Kubernetes Service, or field-programmable gate arrays (FPGA). You retrieve the URI used to access the web service by using the Azure Machine Learning SDK. If authentication is enabled, you can also use the
SDK to get the authentication keys or tokens.
Example:
# URL for the web service
scoring_uri = '<your web service URI>'
# If the service is authenticated, set the key or token
key = '<your key or token>'
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-consume-web-service

Question #77*Topic 3*
HOTSPOT -
You collect data from a nearby weather station. You have a pandas dataframe named weather_df that includes the following data:

Temperature	Observation_time	Humidity	Pressure	Visibility	Days_since_last observation
74	2019/10/2 00:00	0.62	29.87	3	0.5
89	2019/10/2 12:00	0.70	28.88	10	0.5
72	2019/10/3 00:00	0.64	30.00	8	0.5
80	2019/10/3 12:00	0.66	29.75	7	0.5

The data is collected every 12 hours: noon and midnight.
You plan to use automated machine learning to create a time-series model that predicts temperature over the next seven days. For the initial round of training, you want to train a maximum of 50 different models.
You must use the Azure Machine Learning SDK to run an automated machine learning experiment to train these models.
You need to configure the automated machine learning run.
How should you complete the AutoMLConfig definition? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

```
automl_config = AutoMLConfig(task="
```

regression
forecasting
classification
deep learning

```
",
training_data=weather_df,
label_column_name="
```

humidity
pressure
visibility
temperature
days_since_last
observation_time

```
",
time_column_name="
```

humidity
pressure
visibility
temperature
days_since_last
observation_time

```
",
max_horizon=
```

2
6
7
12
14
50

```
,
iterations=
```

2
6
7
12
14
50

```
,
iteration_timeout_minutes=5,
primary_metric="r2_score")
```

Correct Answer:

Answer Area

```
automl_config = AutoMLConfig(task="
```

regression
forecasting
classification
deep learning

",

```
training_data=weather_df,
label_column_name="
```

humidity
pressure
visibility
temperature
days_since_last
observation_time

",

```
time_column_name="
```

humidity
pressure
visibility
temperature
days_since_last
observation_time

",

```
max_horizon=
```

2
6
7
12
14
50

,

```
iterations=
```

2
6
7
12
14
50

,

```
iteration_timeout_minutes=5,
primary_metric="r2_score")
```

Box 1: forcasting -
Task: The type of task to run. Values can be 'classification', 'regression', or 'forecasting' depending on the type of automated ML problem to solve.

Box 2: temperature -
The training data to be used within the experiment. It should contain both training features and a label column (optionally a sample weights column).

Box 3: observation_time -

time_column_name: The name of the time column. This parameter is required when forecasting to specify the datetime column in the input data used for building the time series and inferring its frequency. This setting is being deprecated. Please use forecasting_parameters instead.

Box 4: 7 -
"predicts temperature over the next seven days"
max_horizon: The desired maximum forecast horizon in units of time-series frequency. The default value is 1.
Units are based on the time interval of your training data, e.g., monthly, weekly that the forecaster should predict out. When task type is forecasting, this parameter is required.

Box 5: 50 -
"For the initial round of training, you want to train a maximum of 50 different models."
Iterations: The total number of different algorithm and parameter combinations to test during an automated ML experiment.
Reference:
https://docs.microsoft.com/en-us/python/api/azureml-train-automl-client/azureml.train.automl.automlconfig.automlconfig
Question #78*Topic 3*

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 create a model to forecast weather conditions based on historical data.
You need to create a pipeline that runs a processing script to load data from a datastore and pass the processed data to a machine learning model training script.
Solution: Run the following code:

```
data_store = Datastore.get(ws, “ml-data”)
data_input = DataReference(
    datastore = data_store,
    data_reference_name = “training_data”,
    path_on_datastore = “train/data.txt”)
data_output = PipelineData(“processed_data”, datastore=datastore)
process_step = PythonScriptStep(script_name= “process.py”,
    arguments=[ “- -data”, data_input], outputs=[data_output],
    compute_target=aml_compute, source_directory=process_directory)
train_step = PythonScriptStep(script_name= “train.py”,
      arguments=[“- -data”, data_output], inputs=[data_output],
      compute_target=aml_compute, source_directory=train_directory)
pipeline = Pipeline(workspace=ws, steps = [process_step, train_step])
```

Does the solution meet the goal?

- A. Yes
- B. No

Correct Answer: *A*
The two steps are present: process_step and train_step
Data_input correctly references the data in the data store.
Note:
Data used in pipeline can be produced by one step and consumed in another step by providing a PipelineData object as an output of one step and an input of one or more subsequent steps. PipelineData objects are also used when constructing Pipelines to describe step dependencies. To specify that a step requires the output of another step as input, use a PipelineData object in the constructor of both steps.
For example, the pipeline train step depends on the process_step_output output of the pipeline process step: from azureml.pipeline.core import Pipeline, PipelineData from azureml.pipeline.steps import PythonScriptStep datastore = ws.get_default_datastore() process_step_output = PipelineData("processed_data", datastore=datastore) process_step = PythonScriptStep(script_name="process.py", arguments=["--data_for_train", process_step_output], outputs=[process_step_output], compute_target=aml_compute, source_directory=process_directory) train_step = PythonScriptStep(script_name="train.py", arguments=["--data_for_train", process_step_output], inputs=[process_step_output], compute_target=aml_compute, source_directory=train_directory) pipeline = Pipeline(workspace=ws, steps=[process_step, train_step])
Reference:
https://docs.microsoft.com/en-us/python/api/azureml-pipeline-core/azureml.pipeline.core.pipelinedata?view=azure-ml-py

Question #79*Topic 3*

You run an experiment that uses an AutoMLConfig class to define an automated machine learning task with a maximum of ten model training iterations. The task will attempt to find the best performing model based on a metric named accuracy.
You submit the experiment with the following code:

```
from azureml.core.experiment import Experiment
automl_experiment = Experiment(ws, 'automl_experiment')
automl_run = automl_experiment.submit(automl_config, show_output=True)
```

You need to create Python code that returns the best model that is generated by the automated machine learning task.
Which code segment should you use?

- A. best_model = automl_run.get_details()
- B. best_model = automl_run.get_metrics()
- C. best_model = automl_run.get_file_names()[1]
- D. best_model = automl_run.get_output()[1]

Correct Answer: *D*
The get_output method returns the best run and the fitted model.
Reference:
https://notebooks.azure.com/azureml/projects/azureml-getting-started/html/how-to-use-azureml/automated-machine-learning/classification/auto-ml- classification.ipynb

Question #80*Topic 3*

You plan to use the Hyperdrive feature of Azure Machine Learning to determine the optimal hyperparameter values when training a model.
You must use Hyperdrive to try combinations of the following hyperparameter values. You must not apply an early termination policy.
➾ learning_rate: any value between 0.001 and 0.1
➾ batch_size: 16, 32, or 64
You need to configure the sampling method for the Hyperdrive experiment.
Which two sampling methods can you use? Each correct answer is a complete solution.
NOTE: Each correct selection is worth one point.

- A. No sampling
- B. Grid sampling
- C. Bayesian sampling
- D. Random sampling

Correct Answer: *CD*
C: Bayesian sampling is based on the Bayesian optimization algorithm and makes intelligent choices on the hyperparameter values to sample next. It picks the sample based on how the previous samples performed, such that the new sample improves the reported primary metric.
Bayesian sampling does not support any early termination policy
Example:
from azureml.train.hyperdrive import BayesianParameterSampling from azureml.train.hyperdrive import uniform, choice param_sampling = BayesianParameterSampling({
"learning_rate": uniform(0.05, 0.1),
"batch_size": choice(16, 32, 64, 128)
}
)
D: In random sampling, hyperparameter values are randomly selected from the defined search space. Random sampling allows the search space to include both discrete and continuous hyperparameters.
Incorrect Answers:
B: Grid sampling can be used if your hyperparameter space can be defined as a choice among discrete values and if you have sufficient budget to exhaustively search over all values in the defined search space. Additionally, one can use automated early termination of poorly performing runs, which reduces wastage of resources.
Example, the following space has a total of six samples:
from azureml.train.hyperdrive import GridParameterSampling
from azureml.train.hyperdrive import choice
param_sampling = GridParameterSampling({
"num_hidden_layers": choice(1, 2, 3),
"batch_size": choice(16, 32)
}
)
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-tune-hyperparameters

Question #81*Topic 3*

You are training machine learning models in Azure Machine Learning. You use Hyperdrive to tune the hyperparameters.

In previous model training and tuning runs, many models showed similar performance.
You need to select an early termination policy that meets the following requirements:
➾ accounts for the performance of all previous runs when evaluating the current run avoids comparing the current run with only the best performing run to date
▪
Which two early termination policies should you use? Each correct answer presents part of the solution.
NOTE: Each correct selection is worth one point.

- A. Median stopping
- B. Bandit
- C. Default
- D. Truncation selection

Correct Answer: *AD*
The Median Stopping policy computes running averages across all runs and cancels runs whose best performance is worse than the median of the running averages.
If no policy is specified, the hyperparameter tuning service will let all training runs execute to completion.
Incorrect Answers:
B: BanditPolicy defines an early termination policy based on slack criteria, and a frequency and delay interval for evaluation.
The Bandit policy takes the following configuration parameters: slack_factor: The amount of slack allowed with respect to the best performing training run. This factor specifies the slack as a ratio.
D: The Truncation selection policy periodically cancels the given percentage of runs that rank the lowest for their performance on the primary metric. The policy strives for fairness in ranking the runs by accounting for improving model performance with training time. When ranking a relatively young run, the policy uses the corresponding (and earlier) performance of older runs for comparison. Therefore, runs aren't terminated for having a lower performance because they have run for less time than other runs.
Reference:
https://docs.microsoft.com/en-us/python/api/azureml-train-core/azureml.train.hyperdrive.medianstoppingpolicy https://docs.microsoft.com/en-us/python/api/azureml-train-core/azureml.train.hyperdrive.truncationselectionpolicy
https://docs.microsoft.com/en-us/python/api/azureml-train-core/azureml.train.hyperdrive.banditpolicy

Question #82*Topic 3*

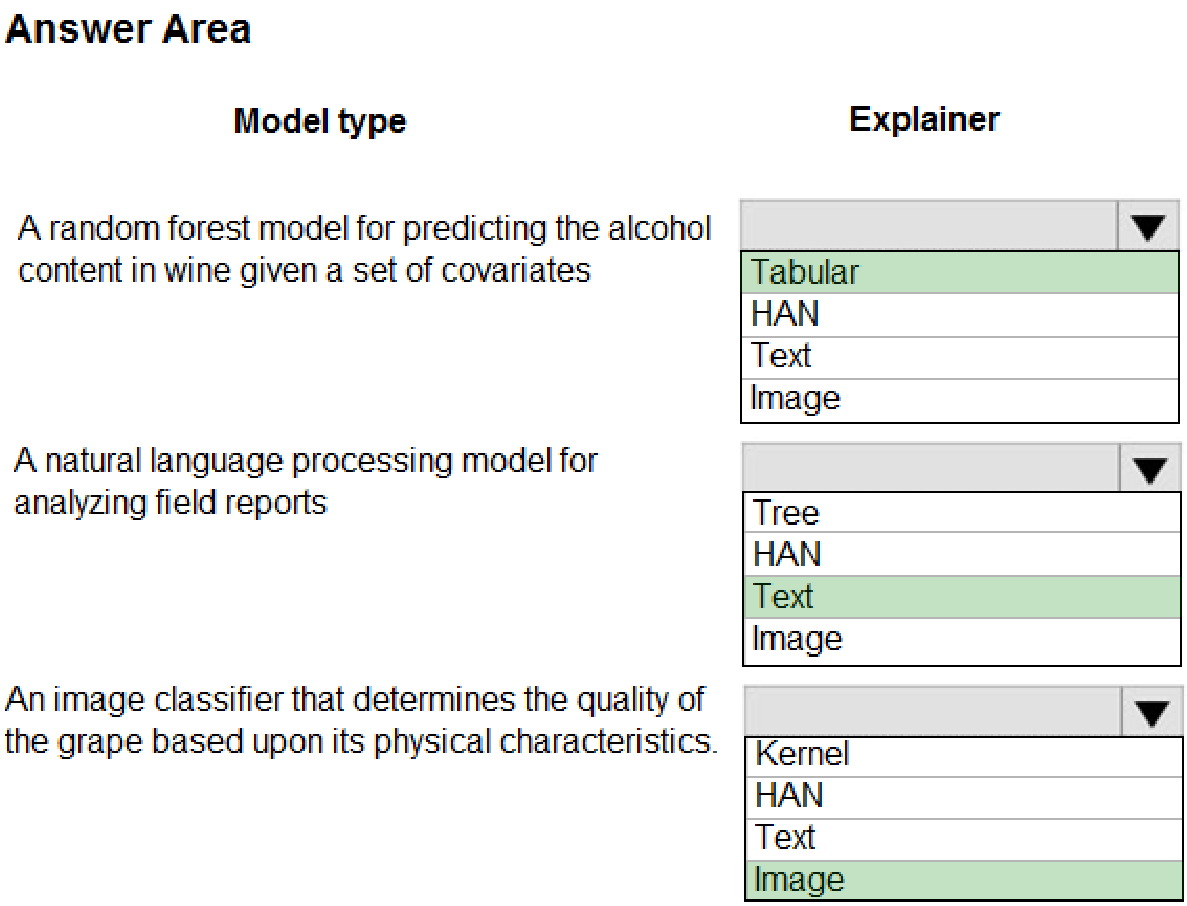
HOTSPOT -
You are hired as a data scientist at a winery. The previous data scientist used Azure Machine Learning.
You need to review the models and explain how each model makes decisions.
Which explainer modules should you use? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Model type	Explainer
A random forest model for predicting the alcohol content in wine given a set of covariates	Tabular HAN Text Image
A natural language processing model for analyzing field reports	Tree HAN Text Image
An image classifier that determines the quality of the grape based upon its physical characteristics.	Kernel HAN Text Image

Correct Answer:



Meta explainers automatically select a suitable direct explainer and generate the best explanation info based on the given model and data sets. The meta explainers leverage all the libraries (SHAP, LIME, Mimic, etc.) that we have integrated or developed. The following are the meta explainers available in the SDK:
Tabular Explainer: Used with tabular datasets.
Text Explainer: Used with text datasets.
Image Explainer: Used with image datasets.

Box 1: Tabular -

Box 2: Text -

Box 3: Image -
Incorrect Answers:
Hierarchical Attention Network (HAN)
HAN was proposed by Yang et al. in 2016. Key features of HAN that differentiates itself from existing approaches to document classification are (1) it exploits the hierarchical nature of text data and (2) attention mechanism is adapted for document classification.
Reference:
https://medium.com/microsoftazure/automated-and-interpretable-machine-learning-d07975741298
Question #83*Topic 3*

HOTSPOT -
You have a dataset that includes home sales data for a city. The dataset includes the following

columns.

Name	Description
Price	The sales price for the house.
Bedrooms	The number of bedrooms in the house.
Size	The size of the house in square feet.
HasGarage	A binary value indicating whether or not the house has a garage.
HomeType	The category of home, for example, apartment, townhouse, single-family home.

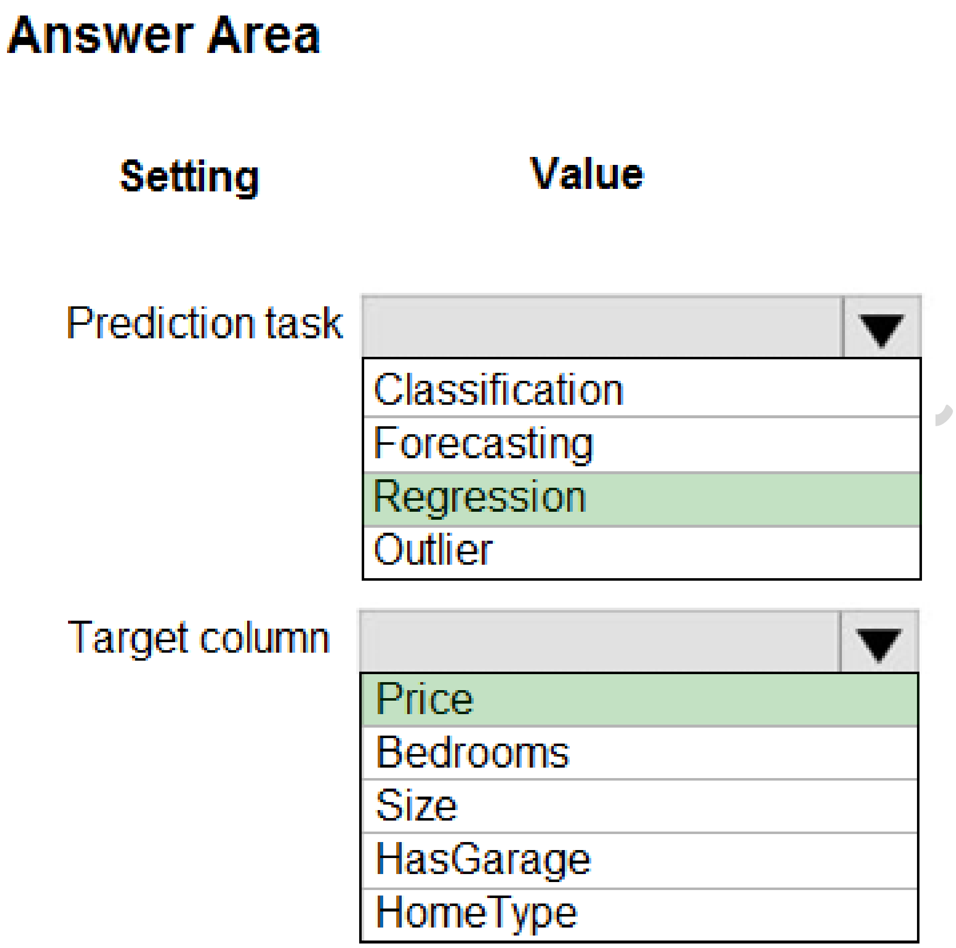
Each row in the dataset corresponds to an individual home sales transaction.
You need to use automated machine learning to generate the best model for predicting the sales price based on the features of the house.
Which values should you use? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.
Hot Area:

Answer Area

Setting **Value**

Prediction task
- Classification
- Forecasting
- Regression
- Outlier

Target column
- Price
- Bedrooms
- Size
- HasGarage
- HomeType



Correct Answer:
Box 1: Regression -
Regression is a supervised machine learning technique used to predict numeric values.

Box 2: Price -
Reference:
https://docs.microsoft.com/en-us/learn/modules/create-regression-model-azure-machine-learning-designer
Question #84*Topic 3*

You use the Azure Machine Learning SDK in a notebook to run an experiment using a script file in an experiment folder.
The experiment fails.
You need to troubleshoot the failed experiment.
What are two possible ways to achieve this goal? Each correct answer presents a complete solution.

- A. Use the get_metrics() method of the run object to retrieve the experiment run logs.
- B. Use the get_details_with_logs() method of the run object to display the experiment run logs.
- C. View the log files for the experiment run in the experiment folder.
- D. View the logs for the experiment run in Azure Machine Learning studio.
- E. Use the get_output() method of the run object to retrieve the experiment run logs.

Correct Answer: *BD*
Use get_details_with_logs() to fetch the run details and logs created by the run.

You can monitor Azure Machine Learning runs and view their logs with the Azure Machine Learning studio.
Incorrect Answers:
A: You can view the metrics of a trained model using run.get_metrics().
E: get_output() gets the output of the step as PipelineData.
Reference:
https://docs.microsoft.com/en-us/python/api/azureml-pipeline-core/azureml.pipeline.core.steprun
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-monitor-view-training-logs
Question #85*Topic 3*

DRAG DROP -
You have an Azure Machine Learning workspace that contains a CPU-based compute cluster and an Azure Kubernetes Service (AKS) inference cluster. You create a tabular dataset containing data that you plan to use to create a classification model.
You need to use the Azure Machine Learning designer to create a web service through which client applications can consume the classification model by submitting new data and getting an immediate prediction as a 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.
Select and Place:

Actions

- Create and run a batch inference pipeline on the compute cluster.
- Deploy a real-time endpoint on the inference cluster.
- Create and run a real-time inference pipeline on the compute cluster.
- Create and run a training pipeline that prepares the data and trains a classification model on the compute cluster.
- Use the automated ML user interface to train a classification model on the compute cluster.
- Create and start a Compute Instance.

Answer Area

Correct Answer:

Actions

- Create and run a batch inference pipeline on the compute cluster.
- Deploy a real-time endpoint on the inference cluster.
- Create and run a real-time inference pipeline on the compute cluster.
- Create and run a training pipeline that prepares the data and trains a classification model on the compute cluster.
- Use the automated ML user interface to train a classification model on the compute cluster.
- Create and start a Compute Instance.

Answer Area

- Create and start a Compute Instance.
- Create and run a training pipeline that prepares the data and trains a classification model on the compute cluster.
- Create and run a real-time inference pipeline on the compute cluster.

Step 1: Create and start a Compute Instance
To train and deploy models using Azure Machine Learning designer, you need compute on which to run the training process, test the model, and host the model in a deployed service.
There are four kinds of compute resource you can create:
Compute Instances: Development workstations that data scientists can use to work with data and models.
Compute Clusters: Scalable clusters of virtual machines for on-demand processing of experiment code.
Inference Clusters: Deployment targets for predictive services that use your trained models.
Attached Compute: Links to existing Azure compute resources, such as Virtual Machines or Azure Databricks clusters.
Step 2: Create and run a training pipeline..
After you've used data transformations to prepare the data, you can use it to train a machine learning model. Create and run a training pipeline
Step 3: Create and run a real-time inference pipeline
After creating and running a pipeline to train the model, you need a second pipeline that performs the same data transformations for new data, and then uses the trained model to inference (in other words, predict) label values based on its features. This pipeline will form the basis for a predictive service that you can publish for applications to use.
Reference:
https://docs.microsoft.com/en-us/learn/modules/create-classification-model-azure-machine-learning-designer/
Question #86*Topic 3*

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 configure the Tune Model Hyperparameters module.
Which two values 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

Correct Answer: B*D*
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.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/two-class-neural-network

Question #87*Topic 3*

HOTSPOT -
You are running a training experiment on remote compute in Azure Machine Learning.
The experiment is configured to use a conda environment that includes the mlflow and azureml-contrib-run packages.
You must use MLflow as the logging package for tracking metrics generated in the experiment.
You need to complete the script for the experiment.
How should you complete the code? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.
Hot Area:

Answer Area

```
import numpy as np
# Import library to log metrics
```

from azureml.core import Run
import mlflow
import logging

```
# Start logging for this run
```

run = Run.get_context()
mlflow.start_run()
logger = logging.getLogger('Run')

```
reg_rate = 0.01
# Log the reg_rate metric
```

run.log('reg_rate', np.float(reg_rate))
mlflow.log_metric('reg_rate', np.float(reg_rate))
logger.info(np.float(reg_rate))

```
# Stop logging for this run
```

run.complete()
mlflow.end_run()
logger.setLevel(logging.INFO)

Answer Area

```
import numpy as np
# Import library to log metrics
```

- from azureml.core import Run
- import mlflow
- import logging

```
# Start logging for this run
```

- run = Run.get_context()
- mlflow.start_run()
- logger = logging.getLogger('Run')

```
reg_rate = 0.01
# Log the reg_rate metric
```

- run.log('reg_rate', np.float(reg_rate))
- mlflow.log_metric('reg_rate', np.float(reg_rate))
- logger.info(np.float(reg_rate))

```
# Stop logging for this run
```

- run.complete()
- mlflow.end_run()
- logger.setLevel(logging.INFO)

Correct Answer:

Box 1: import mlflow -

Import the mlflow and Workspace classes to access MLflow's tracking URI and configure your workspace.

Box 2: mlflow.start_run()

Set the MLflow experiment name with set_experiment() and start your training run with start_run().

Box 3: mlflow.log_metric(' ..')

Use log_metric() to activate the MLflow logging API and begin logging your training run metrics.

Box 4: mlflow.end_run()

Close the run:

run.endRun()

Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-use-mlflow
Question #88*Topic 3*

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

Correct Answer: *D*
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.
Incorrect Answers:
B: If you are building a clustering model, use Sweep Clustering to automatically determine the optimum number of clusters and other parameters.
C: Entire grid: When you select this option, the module loops over a grid predefined by the system, to try different combinations and identify the best learner. This option is useful for cases where you don't know what the best parameter settings might be and want to try all possible combination of values.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/tune-model-hyperparameters

Question #89*Topic 3*

You are building a recurrent neural network to perform a binary classification.
You review the training loss, validation loss, training accuracy, and validation accuracy for each training epoch.
You need to analyze model performance.
You need to identify whether the classification model is overfitted.
Which of the following is correct?

- A. 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.
- 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 increases while the validation loss decreases when training the model.

Correct Answer: *B*
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.
Reference:
https://machinelearningmastery.com/diagnose-overfitting-underfitting-lstm-models/

Question #90*Topic 3*

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 have a Python script named train.py in a local folder named scripts. The script trains a regression model by using scikit-learn. The script includes code to load a training data file which is also located in the scripts folder.
You must run the script as an Azure ML experiment on a compute cluster named aml-compute.
You need to configure the run to ensure that the environment includes the required packages for model training. You have instantiated a variable named aml- compute that references the target compute cluster.
Solution: Run the following code:

```
from azureml.train.estimator import Estimator
sk_est = Estimator(source_directory='./scripts',
 compute_target=aml-compute,
 entry_script='train.py')
```

Does the solution meet the goal?

- A. Yes
- B. No

Correct Answer: *B*
There is a missing line: conda_packages=['scikit-learn'], which is needed.
Correct example:
sk_est = Estimator(source_directory='./my-sklearn-proj',
script_params=script_params,
compute_target=compute_target,
entry_script='train.py',
conda_packages=['scikit-learn'])
Note:

The Estimator class represents a generic estimator to train data using any supplied framework.
This class is designed for use with machine learning frameworks that do not already have an Azure Machine Learning pre-configured estimator. Pre-configured estimators exist for Chainer, PyTorch,

TensorFlow, and SKLearn.
Example:
from azureml.train.estimator import Estimator
script_params = {
# to mount files referenced by mnist dataset
'--data-folder': ds.as_named_input('mnist').as_mount(),
'--regularization': 0.8
}
Reference:
https://docs.microsoft.com/en-us/python/api/azureml-train-core/azureml.train.estimator.estimator
Question #91*Topic 3*

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

Correct Answer: *ACD*
AD: The algorithm terminates when the centroids stabilize or when a specified number of iterations are completed.
C: A measure of how well the centroids represent the members of their clusters is the residual sum of squares or RSS, the squared distance of each vector from its centroid summed over all vectors. RSS is the objective function and our goal is to minimize it.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/k-means-clustering https://nlp.stanford.edu/IR-book/html/htmledition/k-means-1.html

Question #92*Topic 3*

HOTSPOT -
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.

Hot Area:

Answer Area

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

Correct Answer:

Answer Area

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

Box 1: Automatically adjust weights inversely proportional to class frequencies in the input data
The ג€balancedג€ mode uses the values of y to automatically adjust weights inversely proportional to class frequencies in the input data as n_samples / (n_classes * np.bincount(y)).

Box 2: Penalty parameter -
Parameter: C : float, optional (default=1.0)
Penalty parameter C of the error term.
Reference:
https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html
Question #93*Topic 3*

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)

Correct Answer: *C*
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.
Reference:
https://towardsdatascience.com/language-translation-with-rnns-d84d43b40571

Question #94*Topic 3*

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

Correct Answer: *BC*
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.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio/evaluate-model-performance
Question #95*Topic 3*

You create a script that trains a convolutional neural network model over multiple epochs and logs the validation loss after each epoch. The script includes arguments for batch size and learning rate.
You identify a set of batch size and learning rate values that you want to try.
You need to use Azure Machine Learning to find the combination of batch size and learning rate that results in the model with the lowest validation loss.
What should you do?

- A. Run the script in an experiment based on an AutoMLConfig object
- B. Create a PythonScriptStep object for the script and run it in a pipeline
- C. Use the Automated Machine Learning interface in Azure Machine Learning studio
- D. Run the script in an experiment based on a ScriptRunConfig object
- E. Run the script in an experiment based on a HyperDriveConfig object

Correct Answer: *E*
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-tune-hyperparameters
Question #96*Topic 3*

You use the Azure Machine Learning Python SDK to define a pipeline to train a model.
The data used to train the model is read from a folder in a datastore.
You need to ensure the pipeline runs automatically whenever the data in the folder changes.
What should you do?

- A. Set the regenerate_outputs property of the pipeline to True
- B. Create a ScheduleRecurrance object with a Frequency of auto. Use the object to create a Schedule for the pipeline
- C. Create a PipelineParameter with a default value that references the location where the training data is stored
- D. Create a Schedule for the pipeline. Specify the datastore in the datastore property, and the folder containing the training data in the path_on_datastore property

Correct Answer: *D*
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-trigger-published-pipeline
Question #97*Topic 3*

You plan to run a Python script as an Azure Machine Learning experiment.
The script must read files from a hierarchy of folders. The files will be passed to the script as a dataset argument.
You must specify an appropriate mode for the dataset argument.
Which two modes can you use? Each correct answer presents a complete solution.
NOTE: Each correct selection is worth one point.

- A. to_pandas_dataframe()
- B. as_download()
- C. as_upload()
- D. as_mount()

Correct Answer: *B*
Reference:
https://docs.microsoft.com/en-us/python/api/azureml-core/azureml.data.filedataset?view=azure-ml-py
Question #98*Topic 3*

You create a Python script that runs a training experiment in Azure Machine Learning. The script uses the Azure Machine Learning SDK for Python.
You must add a statement that retrieves the names of the logs and outputs generated by the script.
You need to reference a Python class object from the SDK for the statement.
Which class object should you use?

- A. Run
- B. ScriptRunConfig
- C. Workspace
- D. Experiment

Correct Answer: *A*

A run represents a single trial of an experiment. Runs are used to monitor the asynchronous execution of a trial, log metrics and store output of the trial, and to analyze results and access artifacts generated by the trial.

The run Class get_all_logs method downloads all logs for the run to a directory.

Incorrect Answers:

A: A run represents a single trial of an experiment. Runs are used to monitor the asynchronous execution of a trial, log metrics and store output of the trial, and to analyze results and access artifacts generated by the trial.

B: A ScriptRunConfig packages together the configuration information needed to submit a run in Azure ML, including the script, compute target, environment, and any distributed job-specific configs.

Reference:

https://docs.microsoft.com/en-us/python/api/azureml-core/azureml.core.run(class)

Question #99*Topic 3*

You run a script as an experiment in Azure Machine Learning.

You have a Run object named run that references the experiment run. You must review the log files that were generated during the experiment run.

You need to download the log files to a local folder for review.

Which two code segments can you run to achieve this goal? Each correct answer presents a complete solution.

NOTE: Each correct selection is worth one point.

- A. run.get_details()
- B. run.get_file_names()
- C. run.get_metrics()
- D. run.download_files(output_directory='./runfiles')
- E. run.get_all_logs(destination='./runlogs')

Correct Answer: *DE*

The run Class get_all_logs method downloads all logs for the run to a directory.

The run Class get_details gets the definition, status information, current log files, and other details of the run.

Incorrect Answers:

B: The run get_file_names list the files that are stored in association with the run.

Reference:

https://docs.microsoft.com/en-us/python/api/azureml-core/azureml.core.run(class)

Question #100*Topic 3*

You have the following code. The code prepares an experiment to run a script:

```
from azureml.core import Workspace, Experiment, Run, ScriptRunConfig

ws = Workspace.from_config()
script_config = ScriptRunConfig(source_directory='experiment_files',
                                script='experiment.py')

script_experiment = Experiment(workspace=ws, name='script-experiment')
```

The experiment must be run on local computer using the default environment.

You need to add code to start the experiment and run the script.
Which code segment should you use?

- A. run = script_experiment.start_logging()
- B. run = Run(experiment=script_experiment)
- C. ws.get_run(run_id=experiment.id)
- D. run = script_experiment.submit(config=script_config)

Correct Answer: *D*
The experiment class submit method submits an experiment and return the active created run.
Syntax: submit(config, tags=None, **kwargs)
Reference:
https://docs.microsoft.com/en-us/python/api/azureml-core/azureml.core.experiment.experiment

Question #101*Topic 3*

You use the following code to define the steps for a pipeline: from azureml.core import Workspace, Experiment, Run from azureml.pipeline.core import Pipeline from azureml.pipeline.steps import PythonScriptStep ws = Workspace.from_config()
. . .
step1 = PythonScriptStep(name="step1", ...)
step2 = PythonScriptsStep(name="step2", ...)
pipeline_steps = [step1, step2]
You need to add code to run the steps.
Which two code segments can you use to achieve this goal? Each correct answer presents a complete solution.
NOTE: Each correct selection is worth one point.

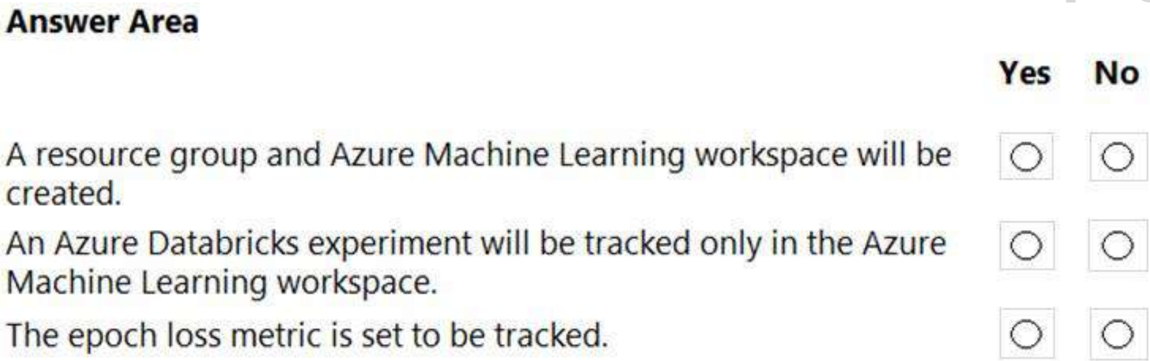
- A. experiment = Experiment(workspace=ws, name='pipeline-experiment') run = experiment.submit(config=pipeline_steps)
- B. run = Run(pipeline_steps)
- C. pipeline = Pipeline(workspace=ws, steps=pipeline_steps) experiment = Experiment(workspace=ws, name='pipeline-experiment') run = experiment.submit(pipeline)
- D. pipeline = Pipeline(workspace=ws, steps=pipeline_steps) run = pipeline.submit(experiment_name='pipeline-experiment')

Correct Answer: *CD*
After you define your steps, you build the pipeline by using some or all of those steps.
# Build the pipeline. Example:
pipeline1 = Pipeline(workspace=ws, steps=[compare_models])
# Submit the pipeline to be run
pipeline_run1 = Experiment(ws, 'Compare_Models_Exp').submit(pipeline1)
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-create-machine-learning-pipelines

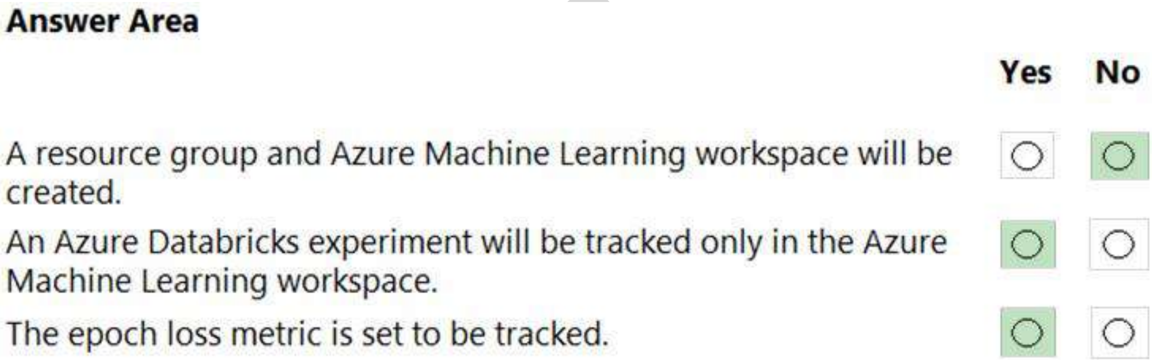
Question #102*Topic 3*

HOTSPOT -
You create an Azure Databricks workspace and a linked Azure Machine Learning workspace.

You have the following Python code segment in the Azure Machine Learning workspace: import mlflow import mlflow.azureml import azureml.mlflow import azureml.core from azureml.core import Workspace subscription_id = 'subscription_id' resourse_group = 'resource_group_name' workspace_name = 'workspace_name' ws = Workspace.get(name=workspace_name, subscription_id=subscription_id, resource_group=resource_group) experimentName = "/Users/{user_name}/{experiment_folder}/{experiment_name}" mlflow.set_experiment(experimentName) uri = ws.get_mlflow_tracking_uri() mlflow.set_tracking_uri(uri)
Instructions: For each of the following statements, select Yes if the statement is true. Otherwise, select No.
NOTE: Each correct selection is worth one point.
Hot Area:



Correct Answer:



Box 1: No -
The Workspace.get method loads an existing workspace without using configuration files. ws = Workspace.get(name="myworkspace", subscription_id='<azure-subscription-id>', resource_group='myresourcegroup')

Box 2: Yes -
MLflow Tracking with Azure Machine Learning lets you store the logged metrics and artifacts from your local runs into your Azure Machine Learning workspace.
The get_mlflow_tracking_uri() method assigns a unique tracking URI address to the workspace, ws, and set_tracking_uri() points the MLflow tracking URI to that address.

Box 3: Yes -
Note: In Deep Learning, epoch means the total dataset is passed forward and backward in a neural network once.

Reference:
https://docs.microsoft.com/en-us/python/api/azureml-core/azureml.core.workspace.workspace
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-use-mlflow

Question #103*Topic 3*
You create and register a model in an Azure Machine Learning workspace.
You must use the Azure Machine Learning SDK to implement a batch inference pipeline that uses a ParallelRunStep to score input data using the model. You must specify a value for the ParallelRunConfig compute_target setting of the pipeline step.
You need to create the compute target.
Which class should you use?

- A. BatchCompute
- B. AdlaCompute
- C. AmlCompute
- D. AksCompute

Correct Answer: *C*
Compute target to use for ParallelRunStep. This parameter may be specified as a compute target object or the string name of a compute target in the workspace.
The compute_target target is of AmlCompute or string.
Note: An Azure Machine Learning Compute (AmlCompute) is a managed-compute infrastructure that allows you to easily create a single or multi-node compute.
The compute is created within your workspace region as a resource that can be shared with other users
Reference:
https://docs.microsoft.com/en-us/python/api/azureml-contrib-pipeline-steps/azureml.contrib.pipeline.steps.parallelrunconfig https://docs.microsoft.com/en-us/python/api/azureml-core/azureml.core.compute.amlcompute(class)

Question #104*Topic 3*

DRAG DROP -
You previously deployed a model that was trained using a tabular dataset named training-dataset, which is based on a folder of CSV files.
Over time, you have collected the features and predicted labels generated by the model in a folder containing a CSV file for each month. You have created two tabular datasets based on the folder containing the inference data: one named predictions-dataset with a schema that matches the training data exactly, including the predicted label; and another named features-dataset with a schema containing all of the feature columns and a timestamp column based on the filename, which includes the day, month, and year.
You need to create a data drift monitor to identify any changing trends in the feature data since the model was trained. To accomplish this, you must define the required datasets for the data drift monitor.
Which datasets should you use to configure the data drift monitor? To answer, drag the appropriate datasets to the correct data drift monitor options. Each source 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.

Select and Place:

Target datasets	Answer Area	
training-dataset	**Baseline dataset**	**Target dataset**
predictions-dataset	Target dataset	Target dataset
features-dataset		

Correct Answer:

Target datasets	Answer Area	
training-dataset	**Baseline dataset**	**Target dataset**
predictions-dataset	training-dataset	predictions-dataset
features-dataset		

Box 1: training-dataset -
Baseline dataset - usually the training dataset for a model.

Box 2: predictions-dataset -
Target dataset - usually model input data - is compared over time to your baseline dataset. This comparison means that your target dataset must have a timestamp column specified.
The monitor will compare the baseline and target datasets.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-monitor-datasets
Question #105*Topic 3*

You plan to run a Python script as an Azure Machine Learning experiment.
The script contains the following code:

```
import os, argparse, glob
from azureml.core import Run

parser = argparse.ArgumentParser()
parser.add_argument('--input-data', type=str, dest='data_folder')
args = parser.parse_args()
data_path = args.data_folder
file_paths = glob.glob(data_path + "/*.jpg")
```

You must specify a file dataset as an input to the script. The dataset consists of multiple large image files and must be streamed directly from its source.
You need to write code to define a ScriptRunConfig object for the experiment and pass the ds dataset as an argument.
Which code segment should you use?

- A. arguments = ['--input-data', ds.to_pandas_dataframe()]
- B. arguments = ['--input-data', ds.as_mount()]
- C. arguments = ['--data-data', ds]

- D. arguments = ['--input-data', ds.as_download()]

Correct Answer: *B*
If you have structured data not yet registered as a dataset, create a TabularDataset and use it directly in your training script for your local or remote experiment.
To load the TabularDataset to pandas DataFrame
df = dataset.to_pandas_dataframe()
Note: TabularDataset represents data in a tabular format created by parsing the provided file or list of files.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-train-with-datasets

Question #106*Topic 3*

You have a Jupyter Notebook that contains Python code that is used to train a model.
You must create a Python script for the production deployment. The solution must minimize code maintenance.
Which two actions should you perform? Each correct answer presents part of the solution.
NOTE: Each correct selection is worth one point.

- A. Refactor the Jupyter Notebook code into functions
- B. Save each function to a separate Python file
- C. Define a main() function in the Python script
- D. Remove all comments and functions from the Python script

Correct Answer: *AC*
C: Python main function is a starting point of any program. When the program is run, the python interpreter runs the code sequentially. Main function is executed only when it is run as a Python program.
A: Refactoring, code style and testing
The first step is to modularise the notebook into a reasonable folder structure, this effectively means to convert files from .ipynb format to .py format, ensure each script has a clear distinct purpose and organise these files in a coherent way.

```
├── src
│   ├── conf            # stores project configurations is json format.
│   ├── main            # main logic for training, predicting and visualisation.
│   ├── resources       # storage of resources such as trained models.
│   ├── template_app    # contains all logic for the flask application.
│   └── utils           # helper functions.
├── tests               # contains projects testing suite.
├── docker-compose.yml  # Docker configurations.
├── Dockerfile          # machine instructions to setup the application and run inside D
├── logs.log            # log files storage.
├── Readme.md
├── requirements.txt    # Python dependancies for installation with pip.
├── run_app.py          # entry point of the project for the Flask application.
└── run.py              # entry point of the project for local usage.
```

Once the project is nicely structured we can tidy up or refactor the code.

Reference:
https://www.guru99.com/learn-python-main-function-with-examples-understand-main.html
https://towardsdatascience.com/from-jupyter-notebook-to-deployment-a-straightforward-example-1838c203a437
Question #107*Topic 3*

HOTSPOT -
You use an Azure Machine Learning workspace.
You create the following Python code:

```
from azureml.core import ScriptRunConfig
src = ScriptRunConfig(source_directory=project_folder,
                      script='train.py'
                      environment=myenv)
```

For each of the following statements, select Yes if the statement is true. Otherwise, select No.
NOTE: Each correct selection is worth one point.
Hot Area:

Answer Area

Statements	Yes	No
The default environment will be created	○	○
The training script will run on local compute	○	○
A script run configuration runs a training script named `train.py` located in a directory defined by the `project_folder` variable	○	○

Correct Answer:

Answer Area

Statements	Yes	No
The default environment will be created	○	○
The training script will run on local compute	○	○
A script run configuration runs a training script named `train.py` located in a directory defined by the `project_folder` variable	○	○

Box 1: No -
Environment is a required parameter. The environment to use for the run. If no environment is specified, azureml.core.runconfig.DEFAULT_CPU_IMAGE will be used as the Docker image for the run.
The following example shows how to instantiate a new environment. from azureml.core import Environment myenv = Environment(name="myenv")

Box 2: Yes -
Parameter compute_target: The compute target where training will happen. This can either be a ComputeTarget object, the name of an existing ComputeTarget, or the string "local". If no compute target is specified, your local machine will be used.

Box 3: Yes -
Parameter source_directory. A local directory containing code files needed for a run.
Parameter script. The file path relative to the source_directory of the script to be run.
Reference:
https://docs.microsoft.com/en-us/python/api/azureml-core/azureml.core.scriptrunconfig
https://docs.microsoft.com/en-us/python/api/azureml-core/azureml.core.environment.environment
Question #108*Topic 3*

HOTSPOT -
You create a Python script named train.py and save it in a folder named scripts. The script uses the scikit-learn framework to train a machine learning model.
You must run the script as an Azure Machine Learning experiment on your local workstation.
You need to write Python code to initiate an experiment that runs the train.py script.
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.
Hot Area:

Answer Area

```
from azureml.core import Experiment, ScriptRunConfig, Environment
from azureml.core.conda_dependencies import CondaDependencies
from azureml.core import Workspace

ws = Workspace.from_config()
py_sk = Environment('sklearn-training')
pkgs = CondaDependencies.create(pip_packages=['scikit-learn', 'azureml-defaults'])
py_sk.python.conda_dependencies = pkgs
script_config = ScriptRunConfig (          ▼ = 'scripts',
                                 script
                                 source_directory
                                 resume_from
                                 arguments
                                           ▼ = 'train.py',
                                 script
                                 arguments
                                 environment
                                 compute_target
                                           ▼ =py_sk)
                                 arguments
                                 resume_from
                                 environment
                                 compute_target

experiment = Experiment(workspace=ws, name='training-experiment')
run = experiment.submit(config=script_config)
```

Correct Answer:

Answer Area

```
from azureml.core import Experiment, ScriptRunConfig, Environment
from azureml.core.conda_dependencies import CondaDependencies
from azureml.core import Workspace

ws = Workspace.from_config()
py_sk = Environment('sklearn-training')
pkgs = CondaDependencies.create(pip_packages=['scikit-learn', 'azureml-defaults'])
py_sk.python.conda_dependencies = pkgs
script_config = ScriptRunConfig ( [▼] = 'scripts',
                                  script
                                  source_directory
                                  resume_from
                                  arguments
                                  [▼] = 'train.py',
                                  script
                                  arguments
                                  environment
                                  compute_target
                                  [▼] =py_sk)
                                  arguments
                                  resume_from
                                  environment
                                  compute_target

experiment = Experiment(workspace=ws, name='training-experiment')
run = experiment.submit(config=script_config)
```

Box 1: source_directory -
source_directory: A local directory containing code files needed for a run.

Box 2: script -
Script: The file path relative to the source_directory of the script to be run.

Box 3: environment -
Reference:
https://docs.microsoft.com/en-us/python/api/azureml-core/azureml.core.scriptrunconfig
Question #109*Topic 3*

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 plan to use a Python script to run an Azure Machine Learning experiment. The script creates a reference to the experiment run context, loads data from a file, identifies the set of unique values for the label column, and completes the experiment run:

```
from azureml.core import Run
import pandas as pd

run = Run.get_context()
data = pd.read_csv('data.csv')
label_vals = data['label'].unique()
# Add code to record metrics here
run.complete()
```

The experiment must record the unique labels in the data as metrics for the run that can be reviewed later.
You must add code to the script to record the unique label values as run metrics at the point indicated by the comment.
Solution: Replace the comment with the following code:
run.log_list('Label Values', label_vals)
Does the solution meet the goal?

- A. Yes
- B. No

Correct Answer: *A*
run.log_list log a list of values to the run with the given name using log_list.
Example: run.log_list("accuracies", [0.6, 0.7, 0.87])
Note:
Data= pd.read_csv('data.csv')
Data is read into a pandas.DataFrame, which is a two-dimensional, size-mutable, potentially heterogeneous tabular data. label_vals =data['label'].unique label_vals contains a list of unique label values.
Reference:
https://www.element61.be/en/resource/azure-machine-learning-services-complete-toolbox-ai
https://docs.microsoft.com/en-us/python/api/azureml-core/azureml.core.run(class)
https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.html

Question #110*Topic 3*

You use the Azure Machine Learning SDK for Python to create a pipeline that includes the following step:

```
step = PythonScriptStep(name= "step1",
                        script_name= "script1.py",
                        compute_target=aml_compute,
                        source_directory=source_directory)
```

The output of the step run must be cached and reused on subsequent runs when the source_directory value has not changed.

You need to define the step.

What should you include in the step definition?

- A. allow_reuse
- B. version
- C. data.as_input(name=…)
- D. hash_paths

Correct Answer: *A*

Question #111*Topic 3*

You are developing a two-step Azure Machine Learning pipeline by using the Azure Machine Learning SDK for Python.

You need to register the output of the pipeline as a new version of a named dataset after the run has been completed.

What should you implement?

- A. the as_input method of the OutputDatasetConfig class
- B. the register_on_complete method of the OutputDatasetConfig class
- C. the as_mount method of the DatasetConsumptionConfig class
- D. the as_download method of the DatasetConsumptionConfig class

Correct Answer: *B*

Question #112*Topic 3*

HOTSPOT

-

You build a data pipeline in an Azure Machine Learning workspace by using the Azure Machine Learning SDK for Python. You create a data preparation step in the data pipeline.

You create the following code fragment in Python:

```
from azureml.core import Dataset
from azureml.pipeline.steps import PythonScriptStep

ds = Dataset.File.from_files([(def_blob_store, 'train-images/')])
ds_input = ds.as_named_input('input1')

source_dir = "./src"
entry_point = "prepare.py"

step = PythonScriptStep(
     script_name=entry_point,
     source_directory=source_dir,
     arguments=["--input", ds_input.as_download(), "--output", output_data1],
     compute_target=compute_target,
     runconfig=aml_run_config,
     allow_reuse=True
)
```

For each of the following statements, select Yes if the statement is true. Otherwise, select No.

NOTE: Each correct selection is worth one point.

Answer Area

Statements	Yes	No
The step will run on the machine defined by the compute_target value, using the configuration stored in the aml_run_config variable.	○	○
A new run will always be generated for this step during pipeline execution.	○	○
Input and output data is logged.	○	○

Correct Answer:

Answer Area

Statements	Yes	No
The step will run on the machine defined by the compute_target value, using the configuration stored in the aml_run_config variable.	$\boxed{\circ}$	○
A new run will always be generated for this step during pipeline execution.	○	$\boxed{\circ}$
Input and output data is logged.	$\boxed{\circ}$	○

Question #113*Topic 3*

HOTSPOT

-

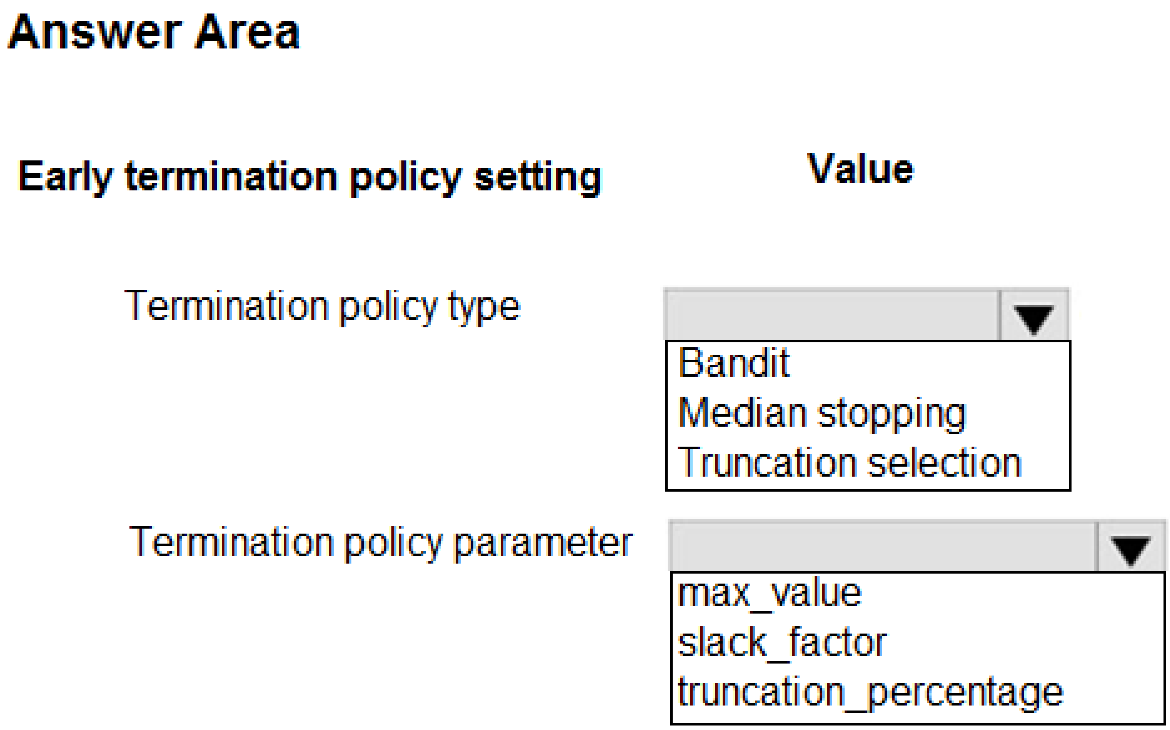
You use Azure Machine Learning to implement hyperparameter tuning.

Training runs must terminate when the primary metric is lowered by 25 percent or more compared to the best performing run.

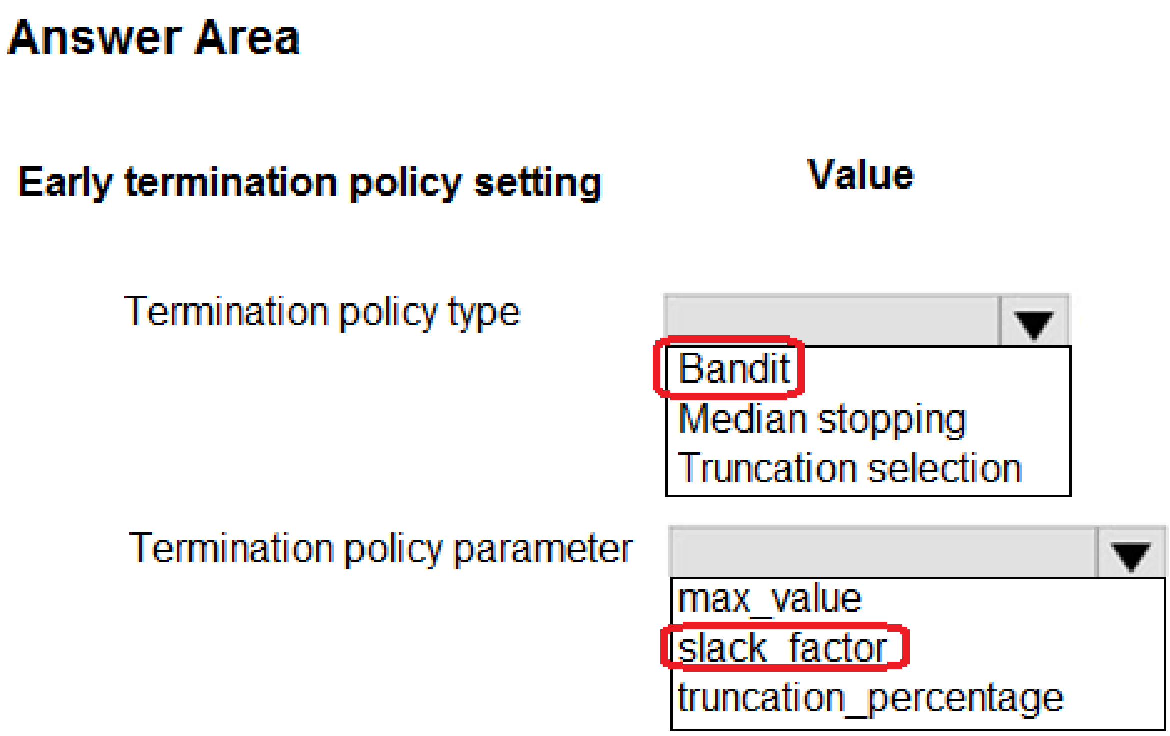
You need to configure an early termination policy to terminate training jobs.

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

NOTE: Each correct selection is worth one point.



Correct Answer:



Question #114*Topic 3*

You are implementing hyperparameter tuning by using Bayesian sampling for a model training from a notebook. The notebook is in an Azure Machine Learning workspace that uses a compute cluster with 20 nodes.

The code implements Bandit termination policy with slack factor set to 0.2 and the HyperDriveConfig class instance with max_concurrent_runs set to 10.

You must increase effectiveness of the tuning process by improving sampling convergence.

You need to select which sampling convergence to use.

What should you select?

- A. Set the value of slack factor of early_termination_policy to 09.
- B. Set the value of max_concurrent_runs of HyperDriveConfig to 4.
- C. Set the value of slack factor of early_termination_policy to 0.1.
- D. Set the value of max_concurrent_runs of HyperDriveConfig to 20.

Correct Answer: *B*

Question #115*Topic 3*

DRAG DROP

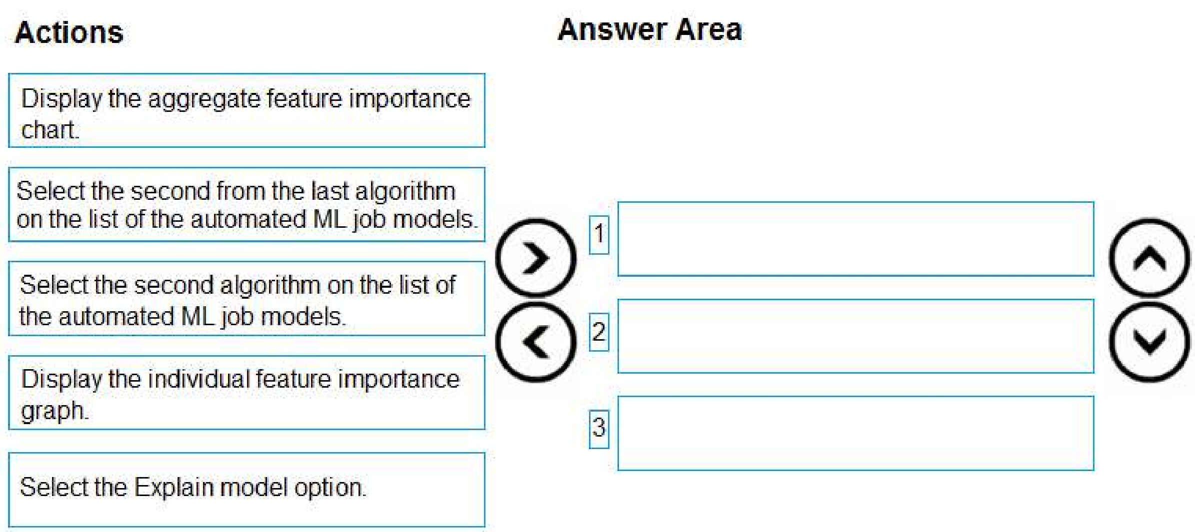
-

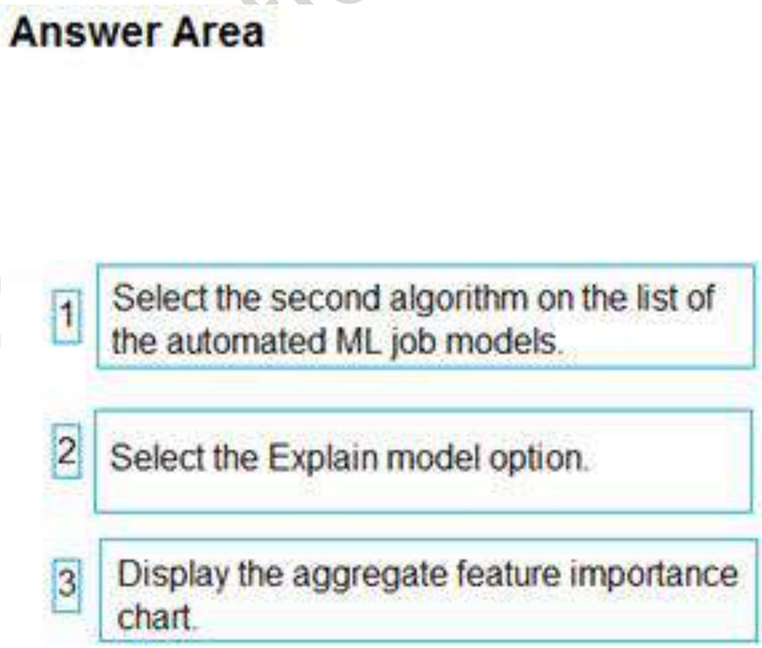
You create an Azure Machine Learning workspace. You are training a classification model with no-code AutoML in Azure Machine Learning studio.

The model must predict if a client of a financial institution will subscribe to a fixed-term deposit. You must identify the feature that has the most influence on the predictions of the model for the second highest scoring algorithm. You must minimize the effort and time to identify the feature.

You need to complete the identification.

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.





Correct Answer:

Question #116*Topic 3*

HOTSPOT

-

You load data from a notebook in an Azure Machine Learning workspace into a pandas dataframe named df. The data contains 10,000 patient records. Each record includes the Age property for the corresponding patient.

You must identify the mean age value from the differentially private data generated by SmartNoise SDK.

You need to complete the Python code that will generate the mean age value from the differentially private data.

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

NOTE: Each correct selection is worth one point.

```
import opendp.smartnoise.core as sn
cols = list(df.columns)
age_range = [0.0, 120.0]
samples = len(df)

with sn. [ ▼ ] as snmethod:
         Analysis()
         QAUILSynthesier()
         MMWEMSynthesier()

data = sn.Dataset(path=data_path, column_names=cols)
    age_dt = sn.to_float(data['Age'])
    age_mean = sn.dp_mean(data = age_dt,
                        privacy_usage = {'[ ▼ ]': .50},
                                         alpha
                                         delta
                                         epsilon
                        data_lower = age_range[0],
                        data_upper = age_range[1],
                        data_rows = samples
                        )
snmethod.release()
print(age_mean.value)
```

Correct Answer:

```
import opendp.smartnoise.core as sn
cols = list(df.columns)
age_range = [0.0, 120.0]
samples = len(df)

with sn. [ ▼ ] as snmethod:
          Analysis()
          QAUILSynthesier()
          MMWEMSynthesier()

data = sn.Dataset(path=data_path, column_names=cols)
    age_dt = sn.to_float(data['Age'])
    age_mean = sn.dp_mean(data = age_dt,
                          privacy_usage = {'[ ▼ ]': .50},
                                            aplha
                                            delta
                                            epsilon
                          data_lower = age_range[0],
                          data_upper = age_range[1],
                          data_rows = samples
                          )
snmethod.release()
print(age_mean.value)
```

Question #117*Topic 3*

HOTSPOT

-

You are developing code to analyze a dataset that includes age information for a large group of diabetes patients. You create an Azure Machine Learning workspace and install all required libraries. You set the privacy budget to 1.0.

You must analyze the dataset and preserve data privacy. The code must run twice before the privacy budget is depleted.

You need to complete the code.

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

NOTE: Each correct selection is worth one point.

Answer Area

```
import pandas as pd
data_path = 'data/diabetes.csv'
diabetes = pd.read_csv(data_path)
import [          ▼] as lib
        opendp.smartnoise.core
        azureml.datadrift
        sklearn.metrics
cols = list(diabetes.columns)

with lib.Analysis() as analysis:
    data = dp.Dataset(path=data_path, column_names=cols)
    age_mean = lib.dp_mean(data = lib.cast(data['age'], type="FLOAT"),
                           privacy_usage = {'[          ▼]': .50},
                           data_lower = 0.,      epsilon
                           data_upper = 100.,    delta
                           data_n = 1000         alpha
                           )
analysis.release()
```

Correct Answer:

Question #118*Topic 3*

You use Azure Machine Learning studio to analyze a dataset containing a decimal column named column1.

You need to verify that the column1 values are normally distributed.

Which statistic should you use?

- A. Max
- B. Type
- C. Profile
- D. Mean

Correct Answer: *C*

Question #119*Topic 3*

HOTSPOT

-

You use Azure Machine Learning to implement hyperparameter tuning with a Bandit early termination policy.

The policy uses a slack_factor set to 0.1. an evaluation interval set to 1, and an evaluation delay set to 5.

You need to evaluate the outcome of the early termination policy.

What should you evaluate? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Answer Area

Scenario	Value
Percentage of worst performing runs to be terminated	1 percent 91 percent 99 percent
Run termination interval	Every interval when metrics are reported, starting at evaluation interval 5. Every 5th interval when metrics are reported. Every 6th interval when metrics are reported.

Correct Answer:

Answer Area

Scenario	Value
Percentage of worst performing runs to be terminated	1 percent 91 percent 99 percent
Run termination interval	Every interval when metrics are reported, starting at evaluation interval 5. Every 5th interval when metrics are reported. Every 6th interval when metrics are reported.

Question #120*Topic 3*

HOTSPOT

-

You train a machine learning model by using Azure Machine Learning.

You use the following training script in Python to log an accuracy value:

```
from azureml.core.run import Run
run_logger = Run.get_context()
run_logger.log("accuracy", float(val_accuracy))
```

You must use a Python script to define a sweep job.

You need to provide the primary metric and goal you want hyperparameter tuning to optimize.

How should you complete the Python script? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Answer Area

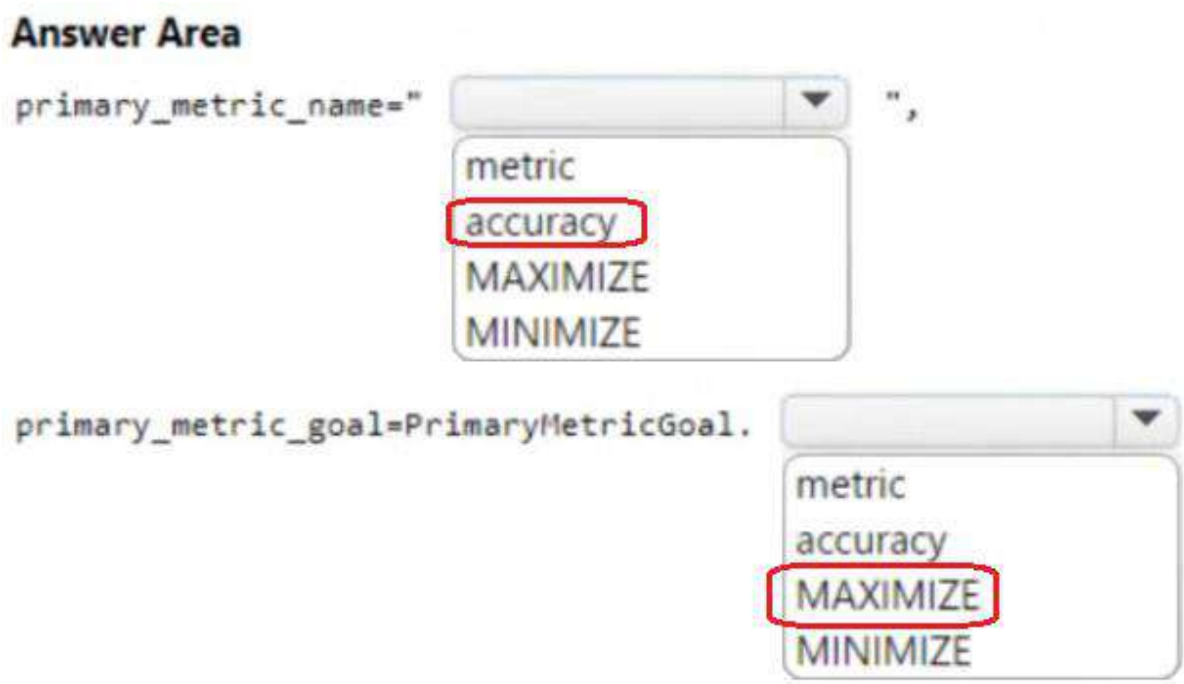
```
primary_metric_name=" [dropdown] ",
```

- metric
- accuracy
- MAXIMIZE
- MINIMIZE

```
primary_metric_goal=PrimaryMetricGoal. [dropdown]
```

- metric
- accuracy
- MAXIMIZE
- MINIMIZE

Correct Answer:



Question #121*Topic 3*

HOTSPOT
-

You have an Azure Machine learning workspace. The workspace contains a dataset with data in a tabular form.

You plan to use the Azure Machine Learning SDK for Python v1 to create a control script that will load the dataset into a pandas dataframe in preparation for model training. The script will accept a parameter designating the dataset.

You need to complete the script.

How should you complete the script? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Answer Area

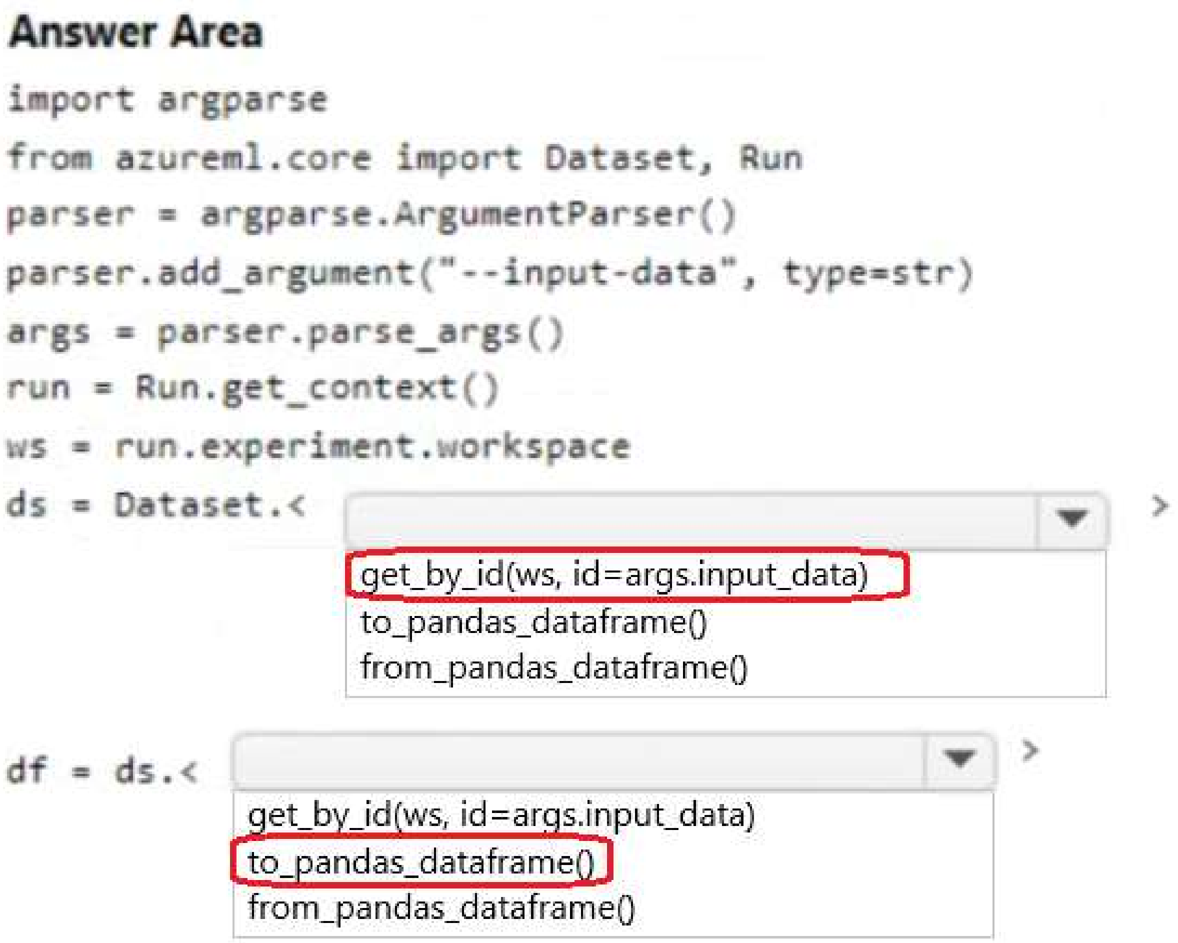
```
import argparse
from azureml.core import Dataset, Run
parser = argparse.ArgumentParser()
parser.add_argument("--input-data", type=str)
args = parser.parse_args()
run = Run.get_context()
ws = run.experiment.workspace
ds = Dataset.<                                    >
```

- get_by_id(ws, id=args.input_data)
- to_pandas_dataframe()
- from_pandas_dataframe()

```
df = ds.<                                    >
```

- get_by_id(ws, id=args.input_data)
- to_pandas_dataframe()
- from_pandas_dataframe()

Correct Answer:



Question #122*Topic 3*

You have a dataset that is stored in an Azure Machine Learning workspace.

You must perform a data analysis for differential privacy by using the SmartNoise SDK.

You need to measure the distribution of reports for repeated queries to ensure that they are balanced.

Which type of test should you perform?

- A. Bias
- B. Privacy
- C. Accuracy
- D. Utility

Correct Answer: *D*

Question #123*Topic 3*

You use the Azure Machine Learning Python SDK to create a batch inference pipeline.

You must publish the batch inference pipeline so that business groups in your organization can use the pipeline. Each business group must be able to specify a different location for the data that the pipeline submits to the model for scoring.

You need to publish the pipeline.

What should you do?

- A. Create multiple endpoints for the published pipeline service and have each business group submit jobs to its own endpoint.
- B. Define a PipelineParameter object for the pipeline and use it to specify the business group-specific input dataset for each pipeline run.
- C. Define a OutputFileDatasetConfig object for the pipeline and use the object to specify the business group-specific input dataset for each pipeline run.
- D. Have each business group run the pipeline on local compute and use a local file for the input data.

Correct Answer: *B*

Question #124*Topic 3*

You create an Azure Machine Learning workspace. You train an MLflow-formatted regression model by using tabular structured data.

You must use a Responsible AI dashboard to access the model.

You need to use the Azure Machine Learning studio UI to generate the Responsible AI dashboard.

What should you do first?

- A. Deploy the model to a managed online endpoint.
- B. Register the model with the workspace.
- C. Create the model explanations.
- D. Convert the model from the MLflow format to a custom format.

Correct Answer: *B*

Question #125*Topic 3*

You are developing a machine learning model by using Azure Machine Learning. You are using multiple text files in tabular format for model data.

You have the following requirements:

• You must use AutoMLjobs to train the model.
• You must use data from specified columns.
• The data concept must support lazy evaluation.

You need to load data into a Pandas dataframe.

Which data concept should you use?

- A. Data asset
- B. URI
- C. Datastore
- D. MLTable

Correct Answer: *D*

Question #126*Topic 3*

You use differential privacy to ensure your reports are private.

The calculated value of the epsilon for your data is 1.8.

You need to modify your data to ensure your reports are private.

Which epsilon value should you accept for your data?

- A. between 0 and 1
- B. between 2 and 3
- C. between 3 and 10
- D. more than 10

Correct Answer: *A*

Question #127*Topic 3*

You create a multi-class image classification model with automated machine learning in Azure Machine Learning.

You need to prepare labeled image data as input for model training in the form of an Azure Machine Learning tabular dataset.

Which data format should you use?

- A. COCO
- B. JSONL
- C. JSON
- D. Pascal VOC

Correct Answer: *B*

Question #128*Topic 3*

You use Azure Machine Learning to train a model.

You must use a sampling method for tuning hyperparameters. The sampling method must pick samples based on how the model performed with previous samples.

You need to select a sampling method.

Which sampling method should you use?

- A. Grid
- B. Bayesian
- C. Random

Correct Answer: *B*

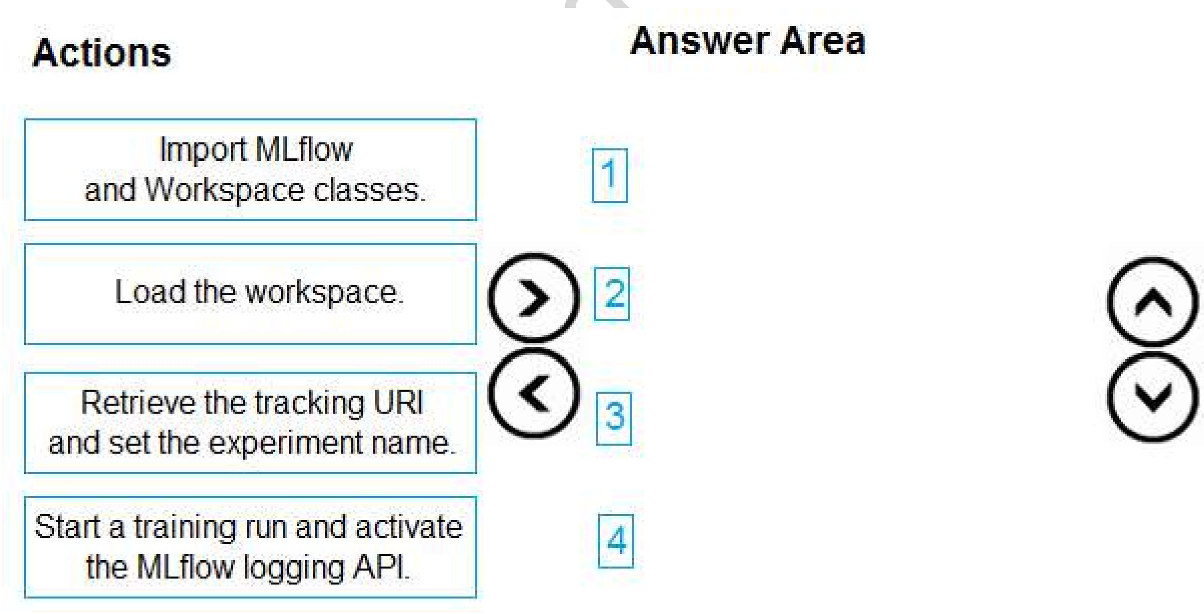
Question #129*Topic 3*

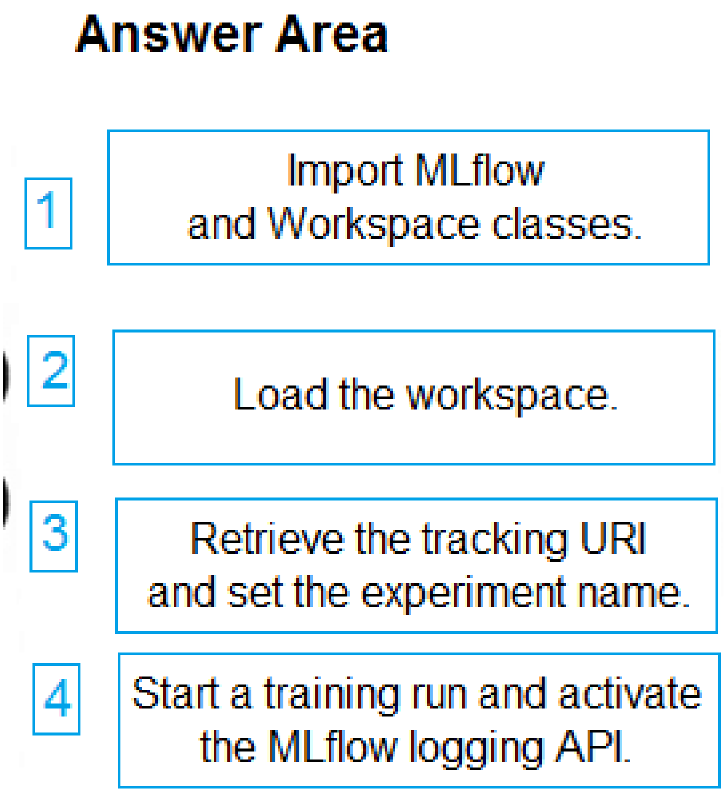
DRAG DROP
-

You have an Azure Machine Learning workspace. You are running an experiment on your local computer.

You need to use MLflow Tracking to store metrics and artifacts from your local experiment runs in the workspace.

In which order should you perform the actions? To answer, move all actions from the list of actions to the answer area and arrange them in the correct order.





Correct Answer:

Question #130*Topic 3*
HOTSPOT
-

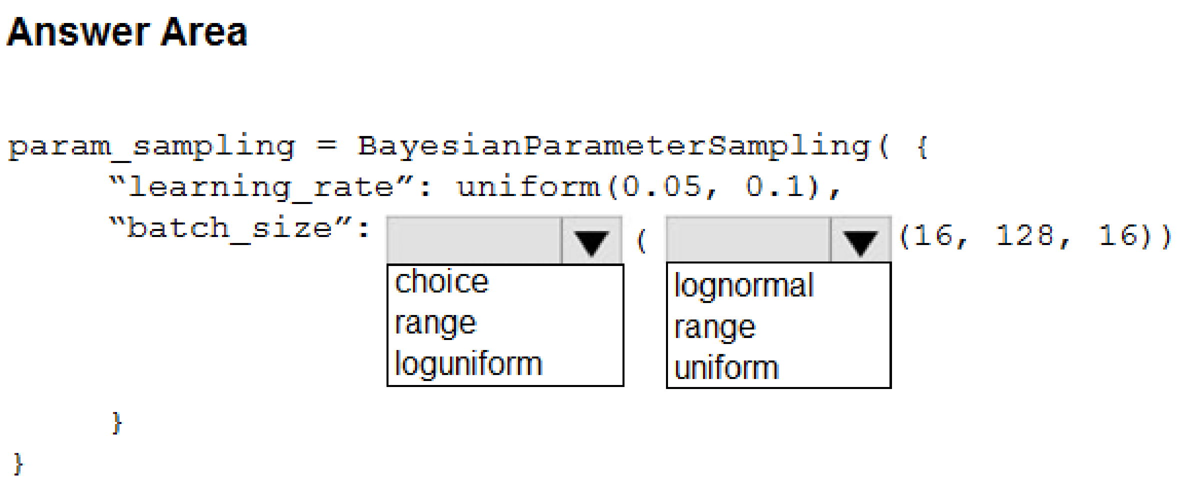
You are implementing hyperparameter tuning for a model training from a notebook. The notebook is in an Azure Machine Learning workspace. You add code that imports all relevant Python libraries.

You must configure Bayesian sampling over the search space for the num_hidden_layers and batch_size hyperparameters.

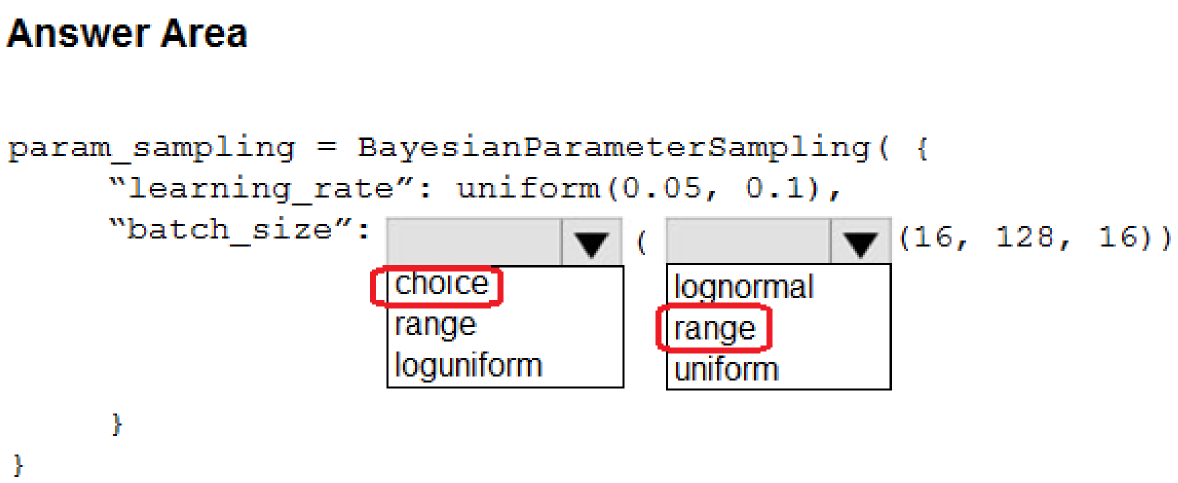
You need to complete the following Python code to configure Bayesian sampling.

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

NOTE: Each correct selection is worth one point.



Correct Answer:



Question #131*Topic 3*

You create a training pipeline by using the Azure Machine Learning designer.

You need to load data into a machine learning pipeline by using the Import Data component.

Which two data sources could you use? Each correct answer presents a complete solution.

NOTE: Each correct selection is worth one point.

- A. Azure SQL Database
- B. Registered dataset
- C. URL via HTTP
- D. Azure Blob storage container through a registered datastore
- E. Azure Data Lake Storage Gen2

Correct Answer: *CD*

Question #132*Topic 3*

HOTSPOT

-

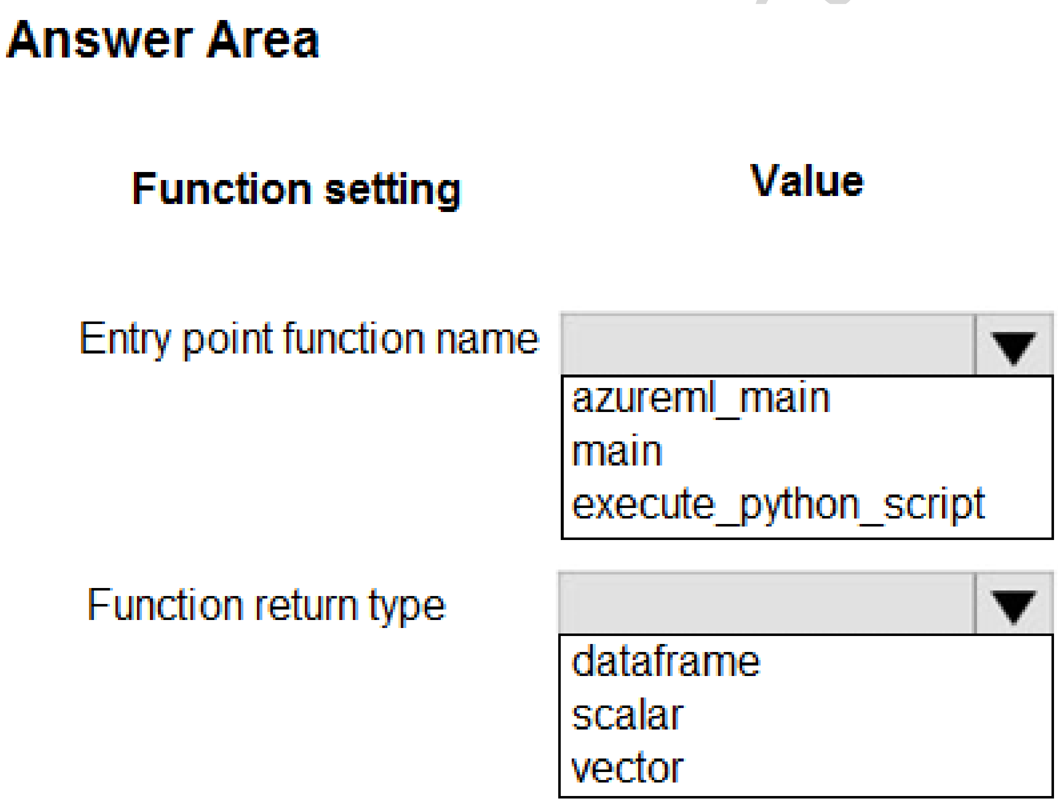
You create an Azure Machine Learning dataset containing automobile price data. The dataset includes 10,000 rows and 10 columns. You use the Azure Machine Learning designer to transform the dataset by using an Execute Python Script component and custom code.

The code must combine three columns to create a new column.

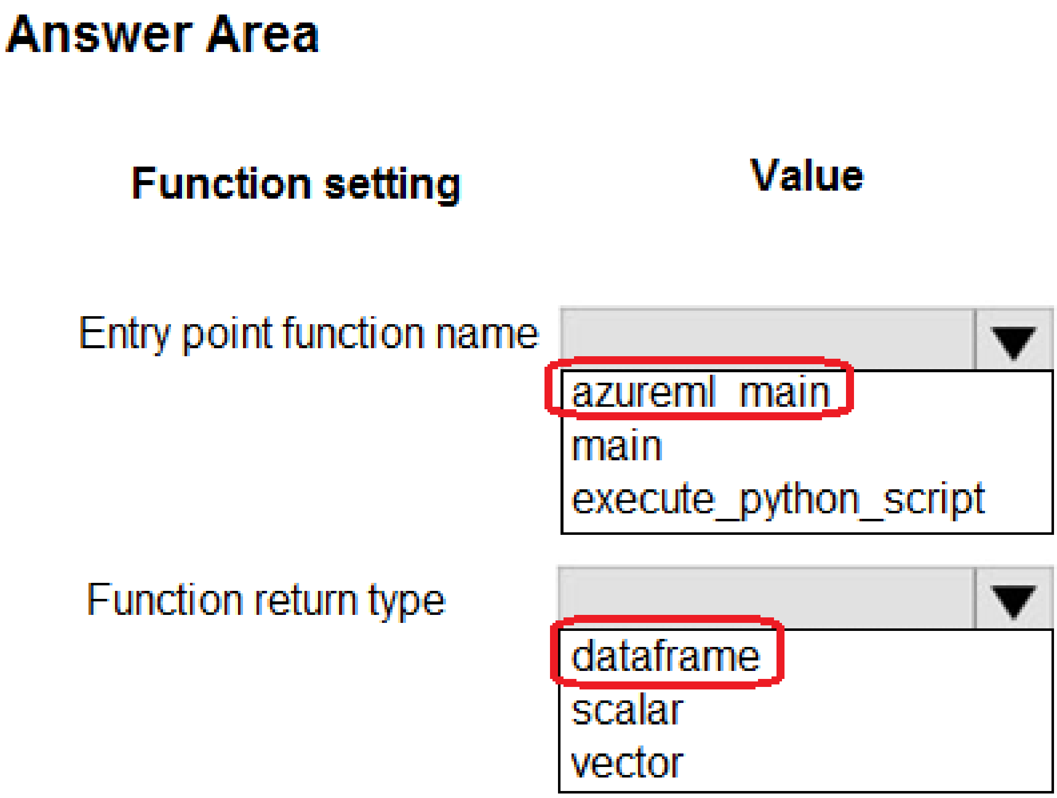
You need to configure the code function.

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

NOTE: Each correct selection is worth one point.



Correct



Answer:

Question #133*Topic 3*

HOTSPOT

-

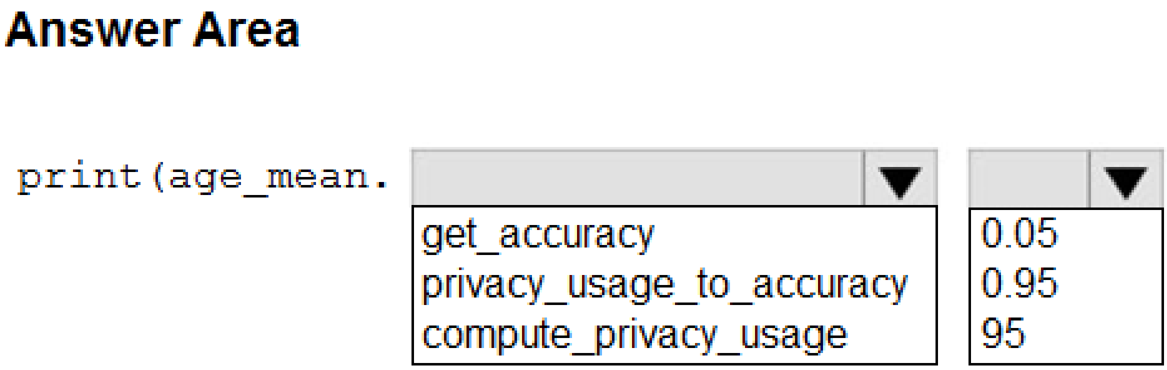
You create an Azure Machine Learning workspace and a dataset. The dataset includes age values for a large group of diabetes patients. You use the dp_mean function from the SmartNoise library to calculate the mean of the age value. You store the value in a variable named age_mean.

You must output the value of the interval rage of released mean values that will be returned 95 percent of the time.

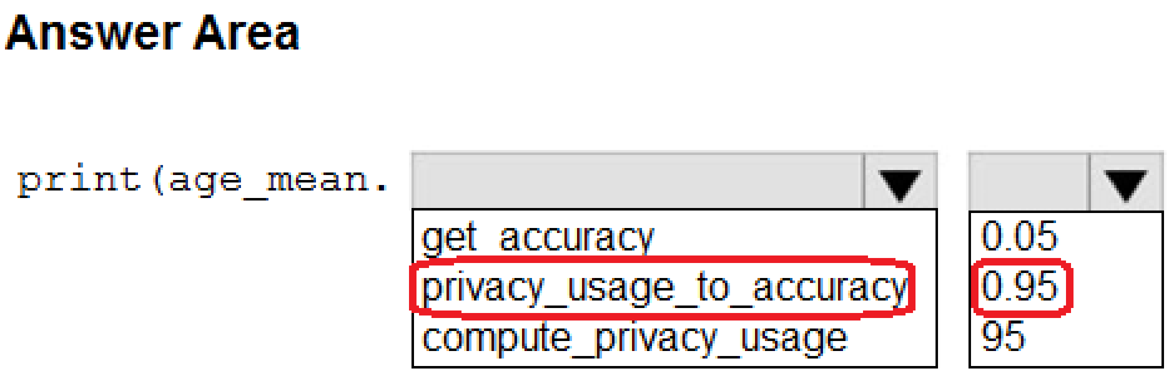
You need to complete the code.

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

NOTE: Each correct selection is worth one point.



Correct Answer:



Question #134*Topic 3*

HOTSPOT

-

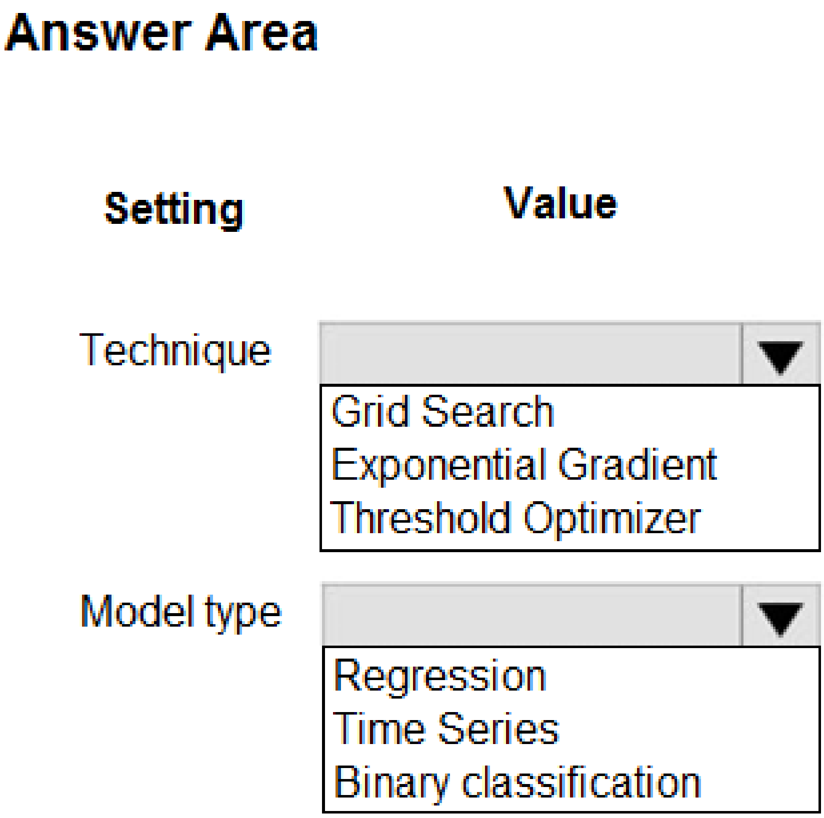
You have machine learning models that produce unfair predictions across sensitive features.

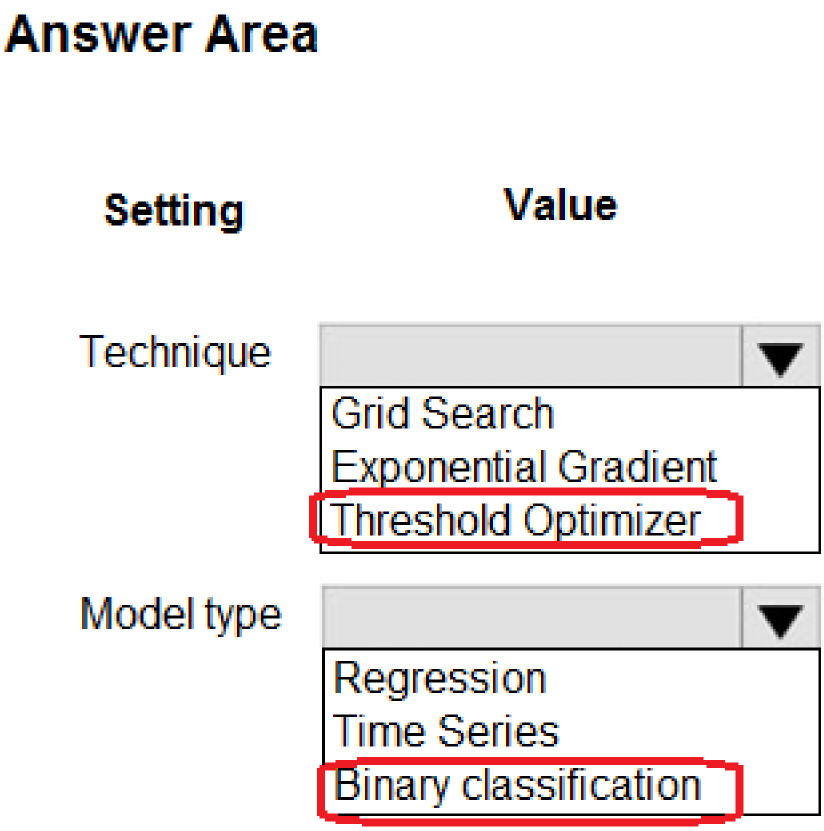
You must use a post-processing technique to apply a constraint to the models to mitigate their unfairness.

You need to select a post-processing technique and model type.

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

NOTE: Each correct selection is worth one point.





Correct Answer:

Question #135*Topic 3*

HOTSPOT

-

You have an Azure Machine Learning workspace.

You plan to use the Azure Machine Learning SDK for Python v1 to submit a job to run a training script.

You need to complete the script to ensure that it will execute the training script.

How should you complete the script? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Answer Area

```
from azureml.core import Workspace, Environment, Experiment, ScriptRunConfig

ws = Workspace.from_config()
env = Environment.get(workspace=ws, name= 'AzureML-Minimal')
exp = Experiment(workspace=ws, name= 'experiment')

src = ScriptRunConfig(source_directory='./src',
         script= 'train.py',
         compute_target= 'compute-cluster'
         environment=env)

run = [▼] . [▼] (config=src)
```

First dropdown: exp / ws / env

Second dropdown: submit / create / write_config

Correct Answer:

Answer Area

```
from azureml.core import Workspace, Environment, Experiment, ScriptRunConfig

ws = Workspace.from_config()
env = Environment.get(workspace=ws, name= 'AzureML-Minimal')
exp = Experiment(workspace=ws, name= 'experiment')

src = ScriptRunConfig(source_directory='./src',
         script= 'train.py',
         compute_target= 'compute-cluster'
         environment=env)

run = [▼] . [▼] (config=src)
```

First dropdown: **exp** / ws / env

Second dropdown: **submit** / create / write_config

Question #136*Topic 3*

HOTSPOT

-

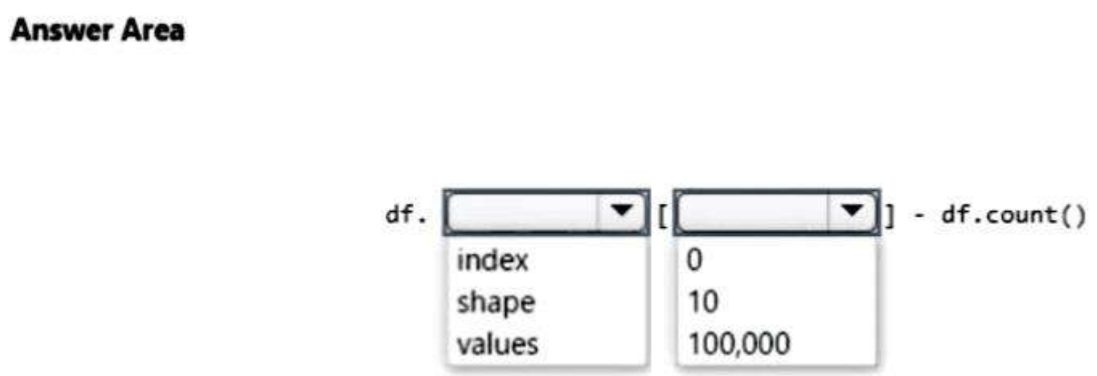
You load data from a notebook in an Azure Machine Learning workspace into a pandas dataframe. The data contains 10,000 records. Each record consists of 10 columns.

You must identify the number of missing values in each of the columns.

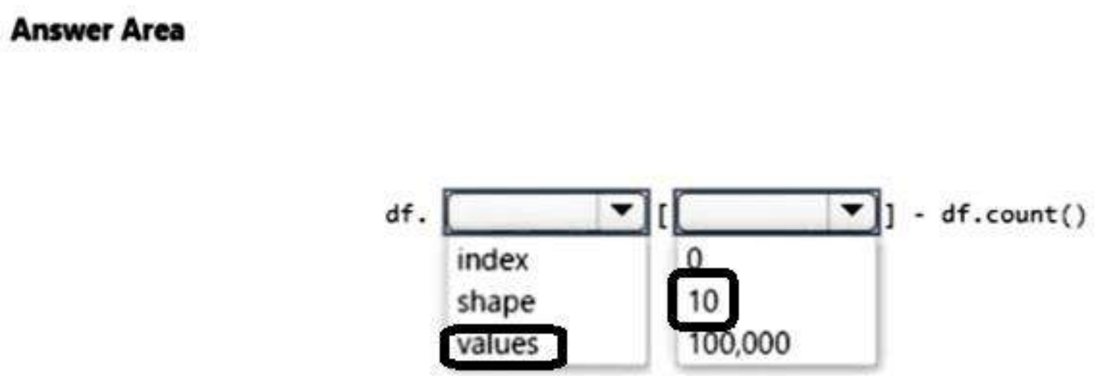
You need to complete the Python code that will return the number of missing values in each of the columns.

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

NOTE: Each correct selection is worth one point.



Correct Answer:



Question #137*Topic 3*

You register a model in an Azure Machine Learning workspace by running the following code:

```
from azureml.core import Model
model = Model.register(workspace=ws,
                       model_name='loan_model',
                       model_path='output/model.pkl')
```

You are creating a scoring script to use in a real-time service for the model.

You need to write code in the scoring script to set the path of the registered model so that it can be loaded by the service. You include the necessary import statements.

Which code segment should you use?

- A. path = Model.get_model_path('loan_model')
- B. path = 'model.pkl'
- C. path = ws.models('loan_model')
- D. path = 'outputs/model.pkl'

Correct Answer: *A*

Question #138*Topic 3*

You are using a ScriptRunConfig object to configure an experiment that uses a script to train a machine learning model.

The script must apply a regularization rate hyperparameter to the algorithm that is used to train the model.

You need to pass the regularization rate in a variable named reg_rate to the script.

Which code segment should you use?

- A.
```
script_config = ScriptRunConfig(source_directory='experiment_files',
                                script='training.py',
                                _telemetry_values = ['--reg_rate', reg_rate],
                                environment=env))
```
- B.
```
script_config = ScriptRunConfig(source_directory='experiment_files',
                                script='training.py',
                                arguments=['--reg_rate', reg_rate],
                                environment=env)
```

- C.

```
script_config = ScriptRunConfig(source_directory='experiment_files',
                                script='training.py',
                                --reg_rate = reg_rate,
                                environment=env))
```

- D.

```
script_config = ScriptRunConfig(source_directory='experiment_files',
                                script='training.py --reg_rate reg_rate',
                                environment=env))
```

Correct Answer: *B*

Question #139*Topic 3*

You are using Azure Machine Learning to monitor a trained and deployed model. You implement Event Grid to respond to Azure Machine Learning events.

Model performance has degraded due to model input data changes.

You need to trigger a remediation ML pipeline based on an Azure Machine Learning event.

Which event should you use?

- A. RunStatusChanged
- B. RunCompleted
- C. DatasetDriftDetected
- D. ModelDeployed

Correct Answer: *C*

Question #140*Topic 3*

HOTSPOT

-

You create an Azure Machine Learning workspace. You train a classification model by using automated machine learning (automated ML) in Azure Machine Learning studio. The training data contains multiple classes that have significantly different numbers of samples.

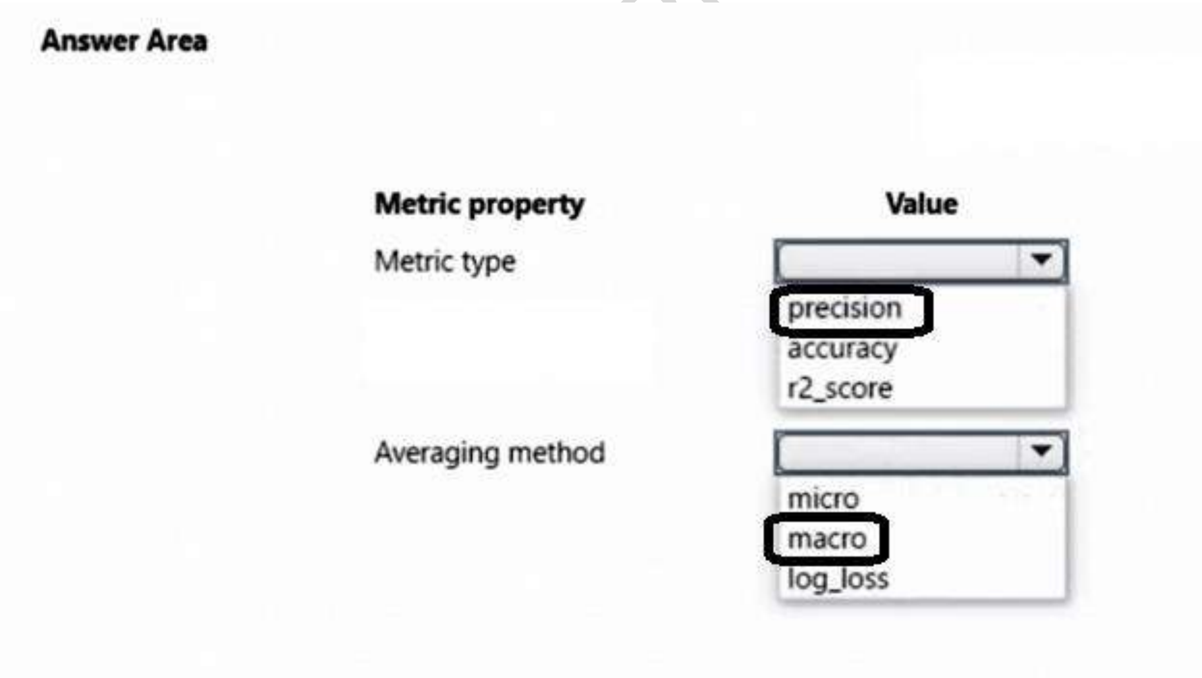
You must use a metric type to avoid labeling negative samples as positive and an averaging method that will minimize the class imbalance.

You need to configure the metric type and the averaging method.

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

NOTE: Each correct selection is worth one point.

Correct Answer:



Question #141*Topic 3*

HOTSPOT

-

You use Azure Machine Learning and SmartNoise Python libraries to implement a differential privacy solution to protect a dataset containing citizen demographics for the city of Seattle in the United

States.

The solution has the following requirements:

• Allow for multiple queries targeting the mean and variance of the citizen's age.
• Ensure full plausible deniability.

You need to define the query rate limit to minimize the risk of re-identification.

What should you configure? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Answer Area

Configuration option

accuracy
noise level
privacy budget

Action

Set the epsilon value to the sum of epsilon values assigned to the mean and variance queries
Set the epsilon value to the larger of the epsilon values assigned to the mean and variance queries
Set the epsilon value to the smaller of the epsilon values assigned to the mean and variance queries

Correct Answer:

Answer Area

Configuration option

accuracy
noise level
privacy budget

Action

Set the epsilon value to the sum of epsilon values assigned to the mean and variance queries
Set the epsilon value to the larger of the epsilon values assigned to the mean and variance queries
Set the epsilon value to the smaller of the epsilon values assigned to the mean and variance queries

Question #142*Topic 3*

You are implementing hyperparameter tuning for a model training from a notebook. The notebook is in an Azure Machine Learning workspace.

You must configure a grid sampling method over the search space for the num_hidden_layers and batch_size hyperparameters.

You need to identify the hyperparameters for the grid sampling.

Which hyperparameter sampling approach should you use?

- A. uniform
- B. qlognormal
- C. choice
- D. normal

Correct Answer: *C*

Question #143*Topic 3*

You create an Azure Machine Learning workspace. You are implementing hyperparameter tuning for a model training from a notebook.

You must configure a Bandit termination policy that provides the following outcome:

If the value of the primary metric of AUC is 0.8 at the point of evaluation intervals, any run with the primary metric value below 0.66 will be terminated.

You need to identify which Bandit termination policy configuration to use.

What should you identify?

- A. Set slack_amount to 0.2.
- B. Set slack_factor to 0.1.
- C. Set slack_factor to 0.2.
- D. Set slack_amount to 0.1.

Correct Answer: *C*

Question #144*Topic 3*

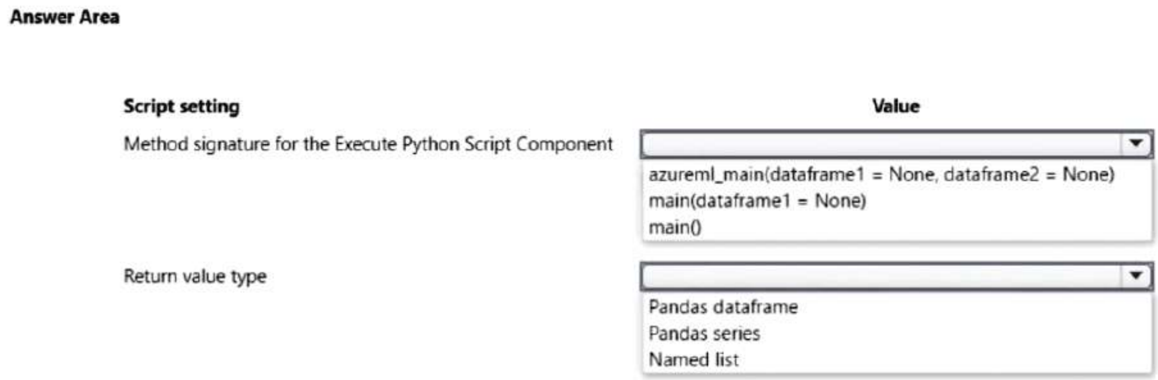
HOTSPOT
-

You are using the Azure Machine Learning designer to transform a dataset by using an Execute Python Script component and custom code.

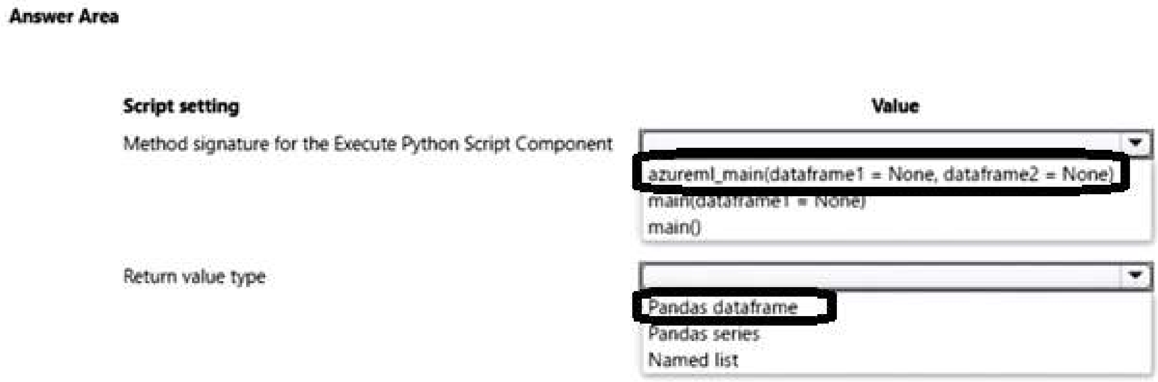
You need to define the method signature for the Execute Python Script component and return value type.

What should you define? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.



Correct Answer:



Question #145*Topic 3*

You need to evaluate the potential risk of exposing personal information based on the values of epsilon and delta for differential privacy. You create a privacy report.

What does an epsilon value greater than one represent?

- A. The privacy of data is preserved and there is limited impact on data accuracy.
- B. There is a high risk of exposing the actual data that is uses to generate the report.
- C. The data used in the report is very noisy.

Correct Answer: *B*

Question #146*Topic 3*

HOTSPOT

-

You manage an Azure Machine Learning workspace named workspace1 by using the Python SDK v2.

The default datastore of workspace1 contains a folder named sample_data. The folder structure contains the following content:

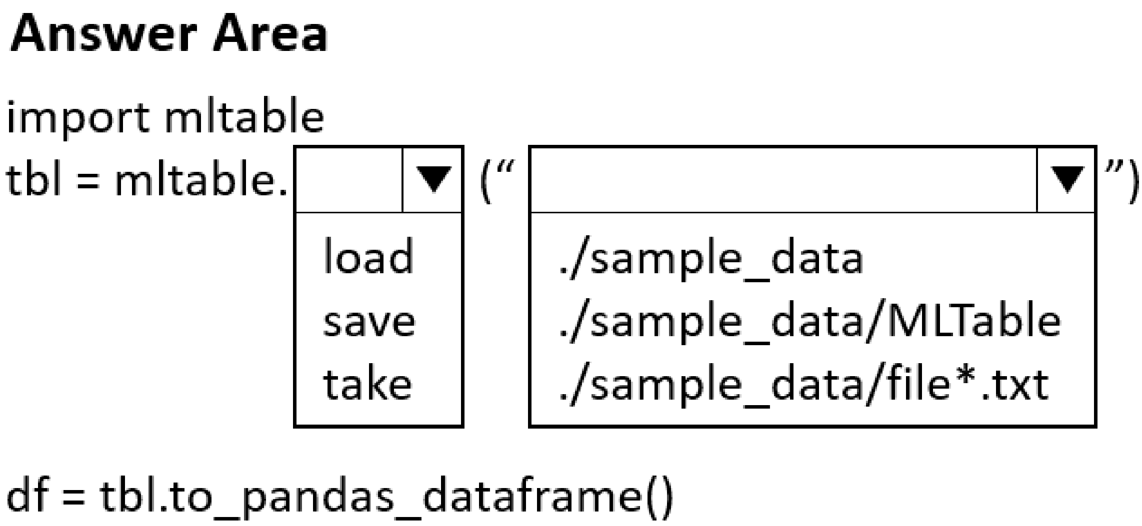
|— sample_data
|— MLTable
|— file1.txt
|— file2.txt
|— file3.txt

You write Python SDK v2 code to materialize the data from the files in the sample_data folder into a Pandas data frame.

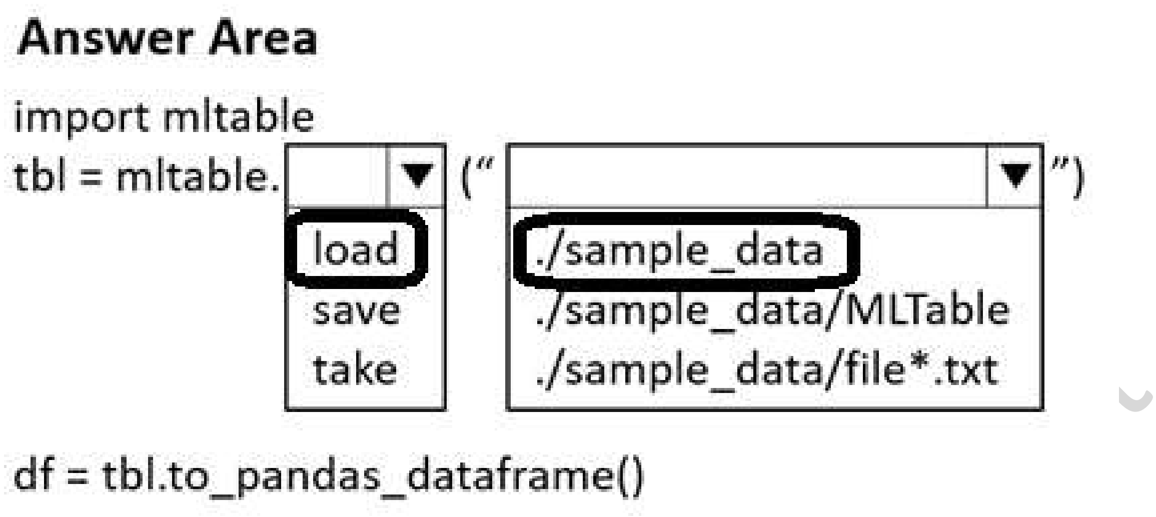
You need to complete the Python SDK v2 code to use the MLTable folder as the materialization blueprint.

How should you complete the code? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.



Correct Answer:



Question #147*Topic 3*

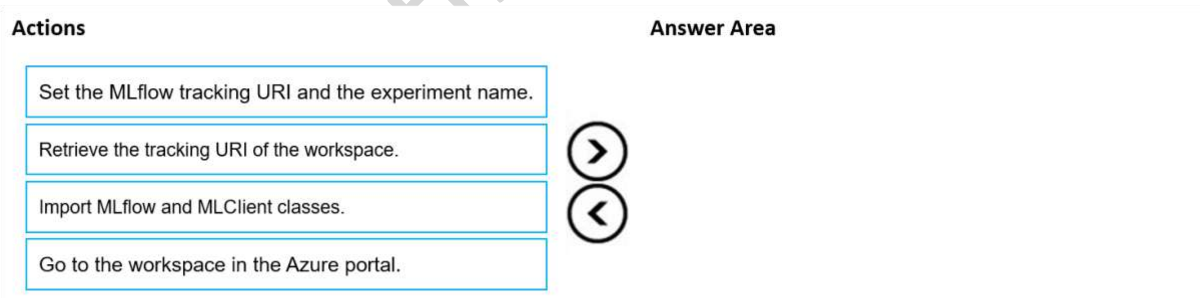
DRAG DROP

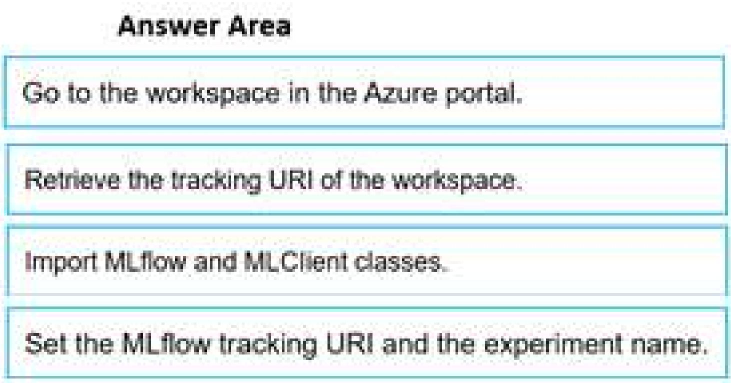
-

You have an Azure Machine Learning workspace. You are running an experiment on your local computer.

You need to ensure that you can use MLflow Tracking with Azure Machine Learning Python SDK v2 to store metrics and artifacts from your local experiment runs m the workspace.

In which order should you perform the actions? To answer, move all actions from the list of actions to the answer area and arrange them in the correct order.




Answer Area

Go to the workspace in the Azure portal.

Retrieve the tracking URI of the workspace.

Import MLflow and MLClient classes.

Set the MLflow tracking URI and the experiment name.

Correct Answer:

Question #148*Topic 3*

You create a workspace by using Azure Machine Learning Studio.

You must run a Python SDK v2 notebook in the workspace by using Azure Machine Learning Studio. You must preserve the current values of variables set in the notebook for the current instance.

You need to maintain the state of the notebook.

What should you do?

- A. Change the compute.
- B. Change the current kernel.
- C. Stop the compute.
- D. Stop the current kernel.

Correct Answer: *D*

Question #149*Topic 3*

HOTSPOT

-

You are using the Azure Machine Learning designer to transform a dataset containing the census data of all nations.

You must use the Split Data component to separate the dataset into two datasets. The first dataset must contain the census data of the United States. The second dataset must include the census data of the remaining nations.

You need to configure the component to create the datasets.

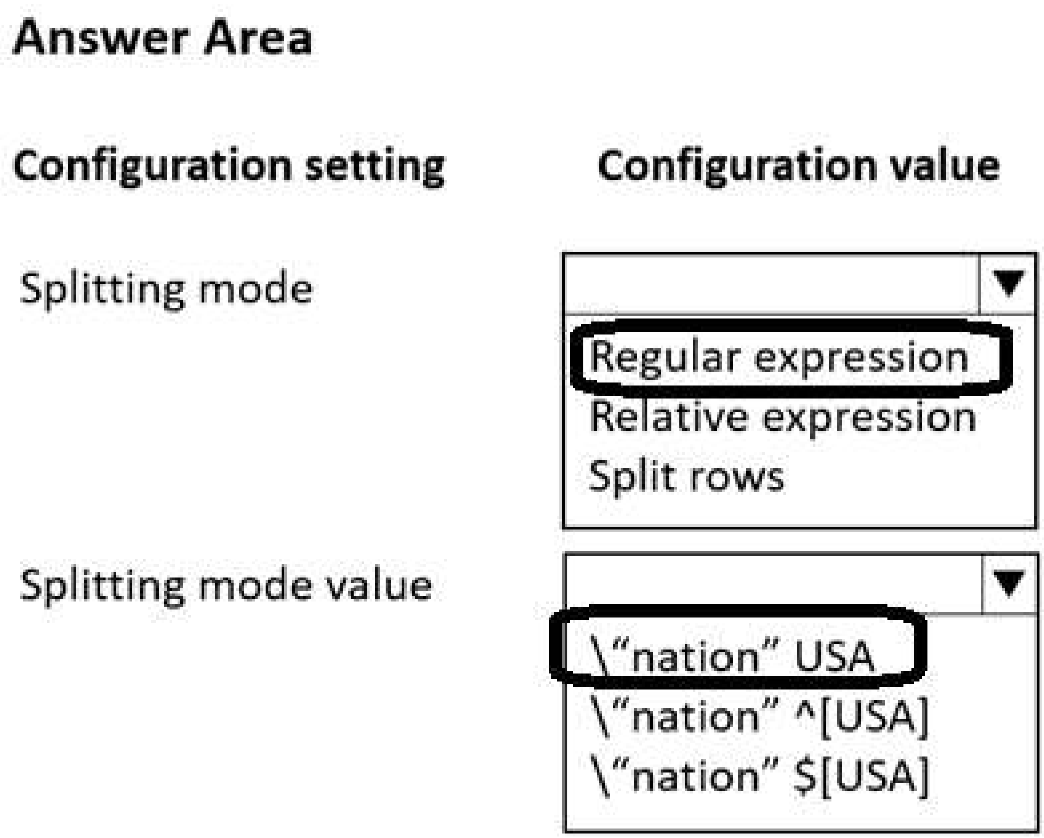
Which configuration values should you use? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Answer Area

Configuration setting	Configuration value
Splitting mode	Regular expression Relative expression Split rows
Splitting mode value	\“nation” USA \“nation” ^[USA] \“nation” $[USA]

Correct Answer:



Question #150*Topic 3*

You create an Azure Machine Learning workspace named workspaces. You create a Python SDK v2 notebook to perform custom model training in workspaces.

You need to run the notebook from Azure Machine Learning Studio in workspaces.

What should you provision first?

- A. default storage account
- B. real-time endpoint
- C. Azure Machine Learning compute cluster
- D. Azure Machine Learning compute instance

Correct Answer: *D*

Question #151*Topic 3*

HOTSPOT

-

You create an Azure Machine Learning workspace.

You must use the Python SDK v2 to implement an experiment from a Jupyter notebook in the workspace. The experiment must log a table in the following format:

table = {
"col1" : [1, 2, 3],
"col2" : [4, 5, 6]
)

You need to complete the Python code to log the table.

How should you complete the code? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

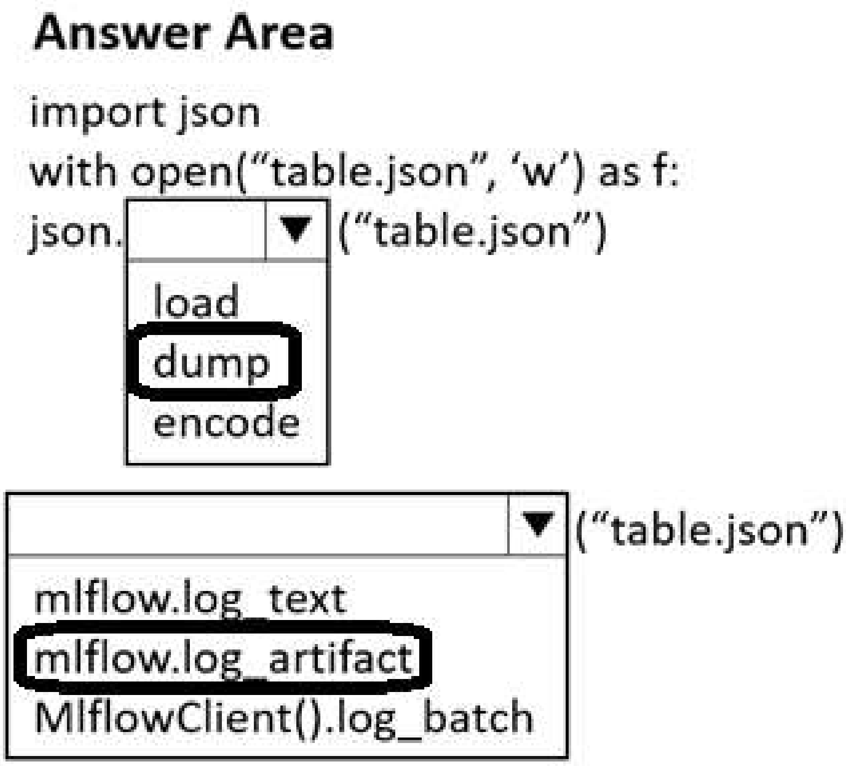
Answer Area

import json
with open(“table.json”, ‘w’) as f:
json.[▼](“table.json”)

- load
- dump
- encode

[▼](“table.json”)

- mlflow.log_text
- mlflow.log_artifact
- MlflowClient().log_batch



Correct Answer:

Question #152*Topic 3*

You plan to run a script as an experiment. The script uses modules from the SciPy library and several Python packages that are not typically installed in a default conda environment.

You plan to run the experiment on your local workstation for small datasets and scale out the experiment by running it on more powerful remote compute dusters for larger datasets.

You need to ensure that the experiment runs successfully on local and remote compute with the least administrative effort.

What should you do?

- A. Leave the environment unspecified for the experiment. Run the expenment by using the default environment.
- B. Create a config.yaml file that defines the required conda packages and save the file in the experiment folder.
- C. Create and register an environment that includes the required packages. Use this environment for all experiment jobs.
- D. Create a virtual machine (VM) by using the required Python configuration and attach the VM as a compute target. Use this compute target for all experiment runs.

Correct Answer: *C*

Question #153*Topic 3*

You have an Azure Machine Learning workspace. You build a deep learning model.

You need to publish a GPU-enabled model as a web service.

Which two compute targets can you use? Each correct answer presents a complete solution.

NOTE: Each correct selection is worth one point.

- A. Azure Kubernetes Service (AKS)
- B. Azure Container Instances (ACI)
- C. Local web service
- D. Azure Machine Learning compute clusters

Correct Answer: *BD*

Question #154*Topic 3*

You create an Azure Machine Learning workspace named workspace1. The workspace contains a Python SDK v2 notebook that uses MLflow to collect model training metrics and artifacts from your local computer.

You must reuse the notebook to run on Azure Machine Learning compute instance in workspace1.

You need to continue to log metrics and artifacts from your data science code.

What should you do?

- A. Instantiate the job class.
- B. Instantiate the MLClient class.
- C. Log in to workspace1.
- D. Configure the tracking URL.

Correct Answer: *D*

Question #155*Topic 3*

You use Azure Machine Learning to train a model.

You must use Bayesian sampling to tune hyperparameters.

You need to select a learning_rate parameter distribution.

Which two distributions can you use? Each correct answer presents a complete solution.

NOTE: Each correct selection is worth one point.

- A. Uniform
- B. Choice
- C. QNormal

- D. Normal
- E. LogUniform

Correct Answer: *AB*

Question #156*Topic 3*

DRAG DROP

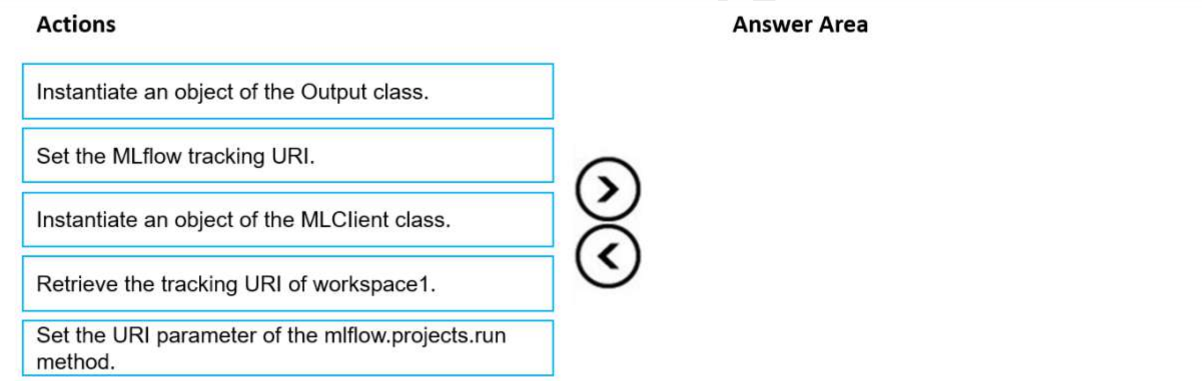
-

You manage an Azure Machine Learning workspace named workspace1 and a Data Science Virtual Machine (DSVM) named DSMV1.

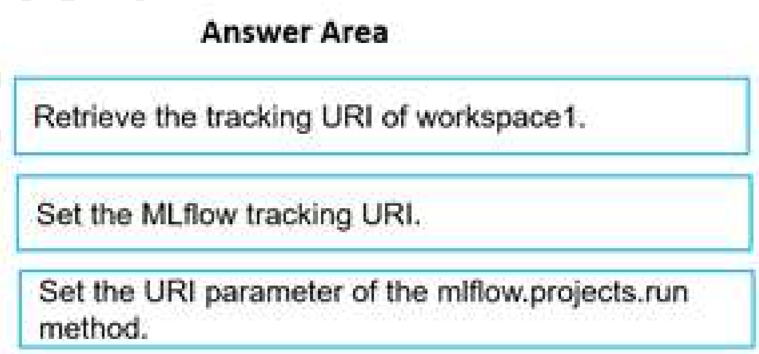
You must run an experiment on DSMV1 by using a Jupyter notebook and Python SDK v2 code. You must store metrics and artifacts in workspace1. You start by creating Python SDK v2 code to import all required packages.

You need to implement the Python SDK v2 code to store metrics and artifacts in workspace1.

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.



Correct Answer:



Question #157*Topic 3*

You create an Azure Machine learning workspace.

You must use the Azure Machine Learning Python SDK v2 to define the search space for discrete hyperparameters. The hyperparameters must consist of a list of predetermined, comma-separated integer values.

You need to import the class from the azure.ai.ml.sweep package used to create the list of values.

Which class should you import?

- A. Choice
- B. Randint
- C. Uniform
- D. Normal

Correct Answer: *B*

Question #158*Topic 3*
HOTSPOT
-

You manage an Azure Machine Learning workspace named workspace1 by using the Python SDK v2. You create a General Purpose v2 Azure storage account named mlstorage1. The storage account includes a publicly accessible container named mlcontainer1.

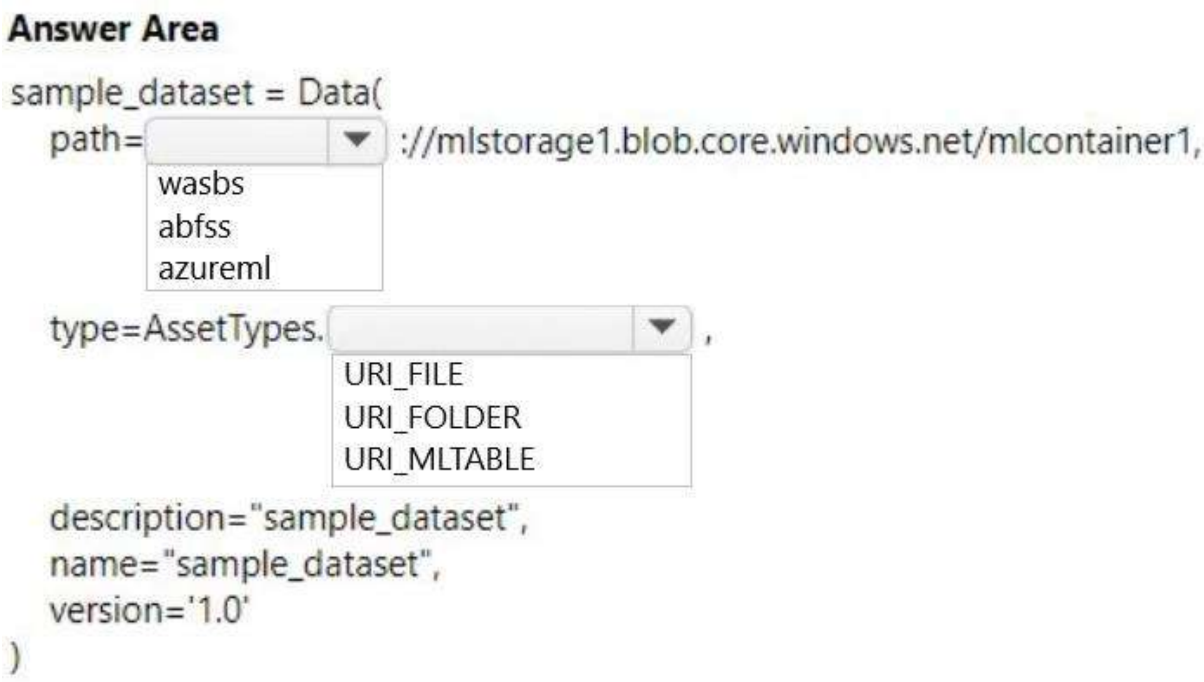
The container stores 10 blobs with files in the CSV format.

You must develop Python SDK v2 code to create a data asset referencing all blobs in the container named mlcontainer1.

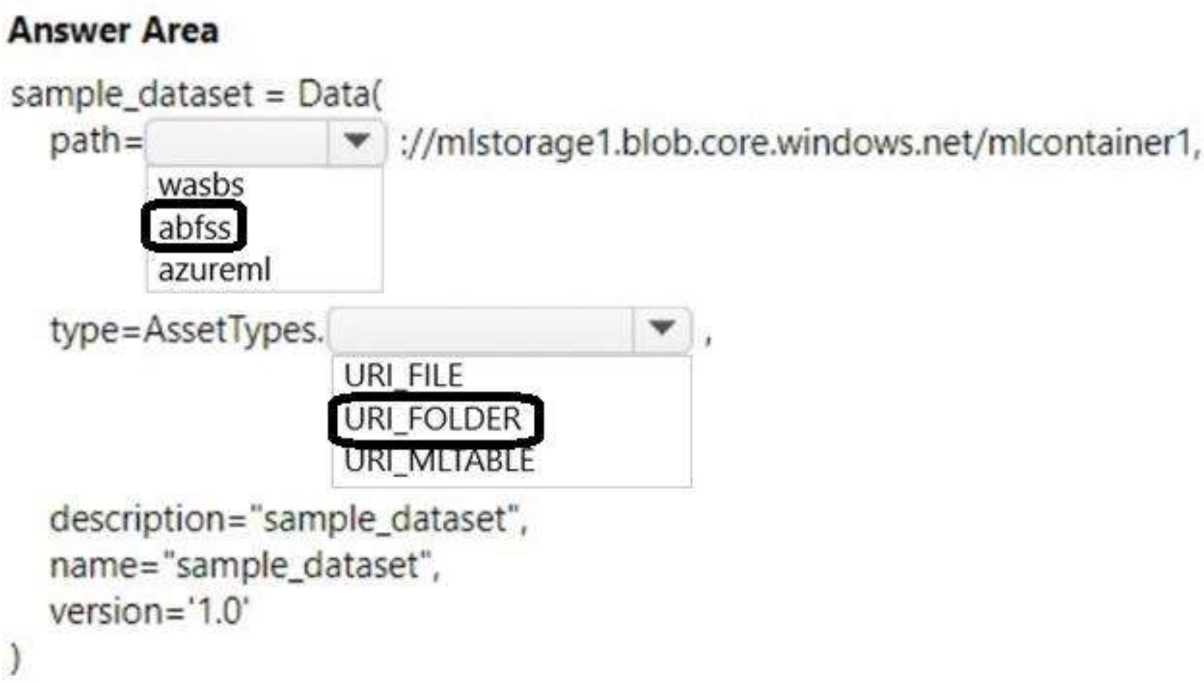
You need to complete the Python SDK v2 code.

How should you complete the code? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.



Correct Answer:



Question #159*Topic 3*

HOTSPOT

-

You monitor an Azure Machine Learning classification training experiment named train.classification on Azure Notebooks.

You must store a table named table as an artifact in Azure Machine Learning Studio during model training.

You need to collect and list the metrics by using MLflow.

How should you complete the code segment? To answer, select the appropriate option in the answer area.

NOTE: Each correct selection is worth one point.

Answer Area

```
from mlflow.tracking import MlflowClient
import json
mlflow.set_experiment("train_classification")
mlflow_run = mlflow.start_run()
row1 = {"table.col1": 5, "table.col2": 10}
mlflow.[            ▼] (row1)
       log_artifact
       log_image
       log_metrics

with open("table.json", 'w') as f:
json.dump(table, f)
mlflow.[            ▼] ("table.json")
       log_artifact
       log_metric
       log_metrics

client = MlflowClient()
finished_mlflow_run =
MlflowClient().get_run([                              ▼] )
                       mlflow_run.info.run_id
                       mlflow_run.info.run_id, path="accuracy_table.json"
                       mlflow_run.info.experiment_id
```

Correct Answer:

Answer Area

```
from mlflow.tracking import MlflowClient
import json
mlflow.set_experiment("train_classification")
mlflow_run = mlflow.start_run()
row1 = {"table.col1": 5, "table.col2": 10}
mlflow.[          ▼] (row1)
```

- log_artifact
- log_image
- log_metrics

```
with open("table.json", 'w') as f:
json.dump(table, f)
mlflow.[          ▼] ("table.json")
```

- log_artifact
- log_metric
- log_metrics

```
client = MlflowClient()
finished_mlflow_run =
MlflowClient().get_run([          ▼] )
```

- mlflow_run.info.run_id
- mlflow_run.info.run_id, path="accuracy_table.json"
- mlflow_run.info.experiment_id

Question #160*Topic 3*

You have a dataset that contains records of patients tested for diabetes. The dataset includes the patient's age.

You plan to create an analysis that will report the mean age value from the differentially private data derived from the dataset.

You need to identify the epsilon value to use in the analysis that minimizes the risk of exposing the actual data.

Which epsilon value should you use?

- A. -1.5
- B. -0.5
- C. 0.5
- D. 1.5

Correct Answer: *C*

Question #161*Topic 3*

HOTSPOT

-

You manage an Azure Machine Learning workspace. You configure an automated machine learning regression training job by using the Azure Machine Learning Python SDK v2.

You configure the regression job by using the following script:

```
regression_job.set_limits(
    timeout_minutes = 60,
    max_concurrent_trials = 5,
    enable_early_termination = True
)
```

For each of the following statements, select Yes if the statement is true. Otherwise, select No.

Answer Area

Statements	Yes	No
The job is terminated if the score is not improving in a specific number of iterations.	○	○
A maximum of five AutoML trials are run in parallel during the regression job.	○	○
One AutoML trial can run for 60 minutes before it is terminated.	○	○
The AutoML trial run can take up to 1 month before it terminates.	○	○

Correct Answer:

Answer Area

Statements	Yes	No
The job is terminated if the score is not improving in a specific number of iterations.	○	☑
A maximum of five AutoML trials are run in parallel during the regression job.	☑	○
One AutoML trial can run for 60 minutes before it is terminated.	☑	○
The AutoML trial run can take up to 1 month before it terminates.	○	☑

Question #162*Topic 3*

DRAG DROP

-

You manage an Azure Machine Learning workspace. You train a model named model1.

You must identify the features to modify for a differing model prediction result.

You need to configure the Responsible AI (RAI) dashboard for model1.

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.

Actions

- Add the error analysis component to the Responsible AI Insights dashboard.
- Use the Gather Responsible AI Insights dashboard component to present the dashboard.
- Add the counterfactuals component to the Responsible AI Insights dashboard.
- Load and configure the Responsible AI Insights dashboard constructor component.
- Add the explanation component to the Responsible AI Insights dashboard.
- Add the causal component to the Responsible AI Insights dashboard.

Answer Area

Answer Area

1. Load and configure the Responsible AI Insights dashboard constructor component.
2. Add the counterfactuals component to the Responsible AI Insights dashboard.
3. Use the Gather Responsible AI Insights dashboard component to present the dashboard.

Correct Answer:

Question #163*Topic 3*

HOTSPOT

\-

You are using hyperparameter tuning in Azure Machine Learning Python SDK v2 to train a model.

You configure the hyperparameter tuning experiment by running the following code:

```
command_job_for_sweep = command_job(
    learning_rate=Normal(10, 3),
    keep_probability=Uniform(0.05, 0.1),
    batch_size=Choice(values=[16, 32, 64, 128]),
    number_of_hidden_layers=Choice(range(3,5))
)
```

For each of the following statements, select Yes if the statement is true. Otherwise, select No.

NOTE: Each correct selection is worth one point.

Answer Area

Statements	Yes	No
By defining sampling in this manner, every possible combination of the parameters will be tested.	○	○
Random values of the learning_rate parameter will be selected from a normal distribution with a mean of 10 and a standard deviation of 3.	○	○
The keep_probability parameter value will always be either **0.05** or **0.1**.	○	○
Random values for the number_of_hidden_layers parameter will be selected from a normal distribution with a mean of 3 and a standard deviation of 5.	○	○

Correct Answer:

Answer Area

Statements	Yes	No
By defining sampling in this manner, every possible combination of the parameters will be tested.	$\boxed{○}$	○
Random values of the learning_rate parameter will be selected from a normal distribution with a mean of 10 and a standard deviation of 3.	$\boxed{○}$	○
The keep_probability parameter value will always be either **0.05** or **0.1**.	○	$\boxed{○}$
Random values for the number_of_hidden_layers parameter will be selected from a normal distribution with a mean of 3 and a standard deviation of 5.	○	$\boxed{○}$

Question #164*Topic 3*

HOTSPOT

-

You use Azure Machine Learning to train a machine learning model.

You use the following training script in Python to perform logging:

import mlflow
mlflow.log_metric("accuracy", float(vel_accuracy))

You must use a Python script to define a sweep job.

You need to provide the primary metric and goal you want hyperparameter tuning to optimize.

How should you complete the Python script? To answer, select the appropriate options in the answer area.

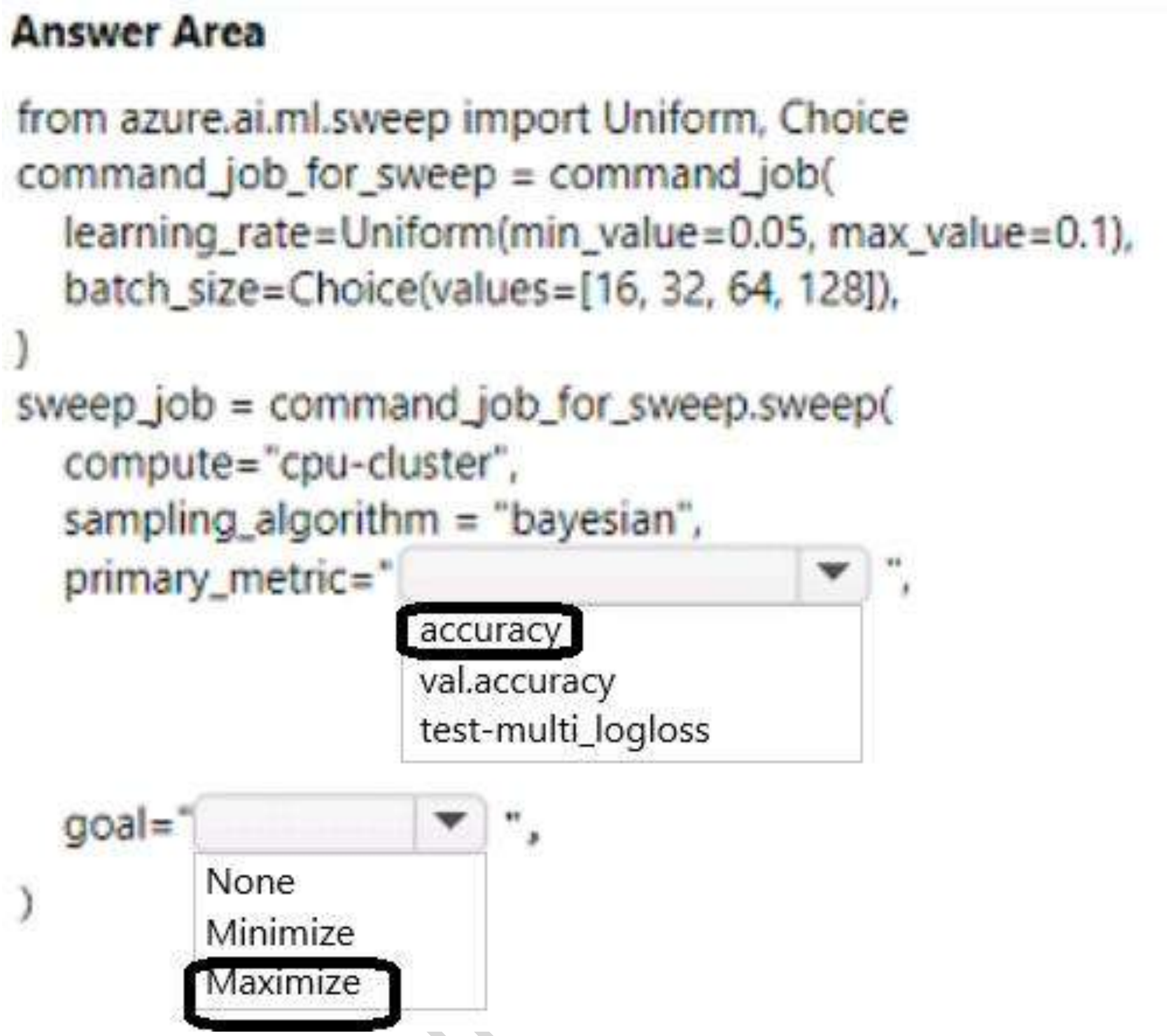
NOTE: Each correct selection is worth one point.

Answer Area

```
from azure.ai.ml.sweep import Uniform, Choice
command_job_for_sweep = command_job(
    learning_rate=Uniform(min_value=0.05, max_value=0.1),
    batch_size=Choice(values=[16, 32, 64, 128]),
)
sweep_job = command_job_for_sweep.sweep(
    compute="cpu-cluster",
    sampling_algorithm = "bayesian",
    primary_metric="[ ▼ ]",
                    accuracy
                    val.accuracy
                    test-multi_logloss

    goal="[ ▼ ]",
          None
          Minimize
          Maximize
)
```


Correct Answer:



Question #165*Topic 3*

You create a workspace by using Azure Machine Learning Studio.

You must run a Python SDK v2 notebook in the workspace by using Azure Machine Learning Studio.

You need to reset the state of the notebook.

Which three actions should you use? Each correct answer presents a complete solution.

NOTE: Each correct selection is worth one point.

- A. Stop the current kernel.
- B. Change the compute.
- C. Reset the compute.
- D. Navigate to another section of the workspace.
- E. Change the current kernel.

Correct Answer: *ADE*

Question #166*Topic 3*

You manage an Azure Machine Learning workspace.

You must log multiple metrics by using MLflow.

You need to maximize logging performance.

What are two possible ways to achieve this goal? Each correct answer presents a complete solution.

NOTE: Each correct selection is worth one point.

- A. MLflowClient.log_batch
- B. mlflow.log_metrics
- C. mlflow.log_metric
- D. mlflow.log_param

Correct Answer: *AC*

Question #167*Topic 3*

HOTSPOT

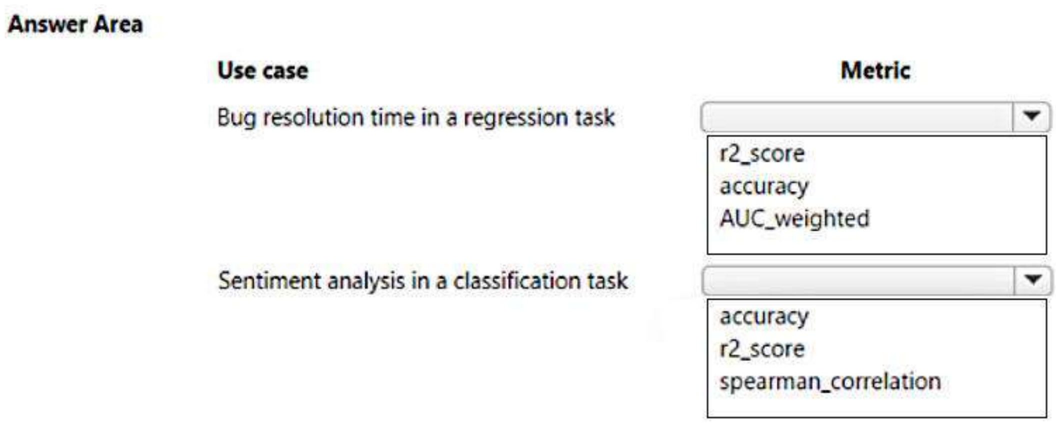
-

You create multiple machine learning models by using automated machine learning.

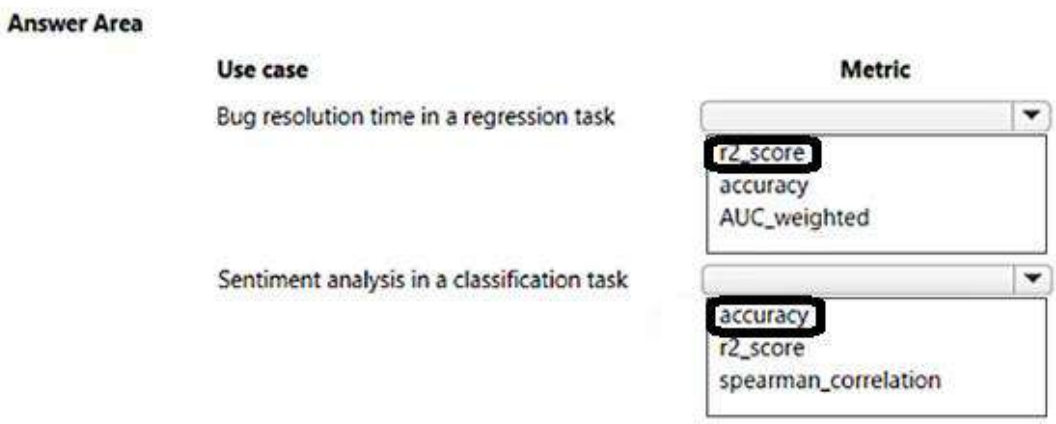
You need to configure a primary metric for each use case.

Which metrics should you configure? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.



Correct Answer:



Question #168*Topic 3*

HOTSPOT

-

You are running a training experiment on remote compute in Azure Machine Learning (ML) by using Azure ML SDK v2 for Python.

The experiment is configured to use a conda environment that includes all required packages.

You must track metrics generated in the experiment.

You need to complete the script for the experiment.

How should you complete the code? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Answer Area

```
import numpy as np
```

- import os
- import mlflow
- import random

- os.system(f"random()")
- mlflow.start_run()
- print(random())

```
reg_rate = 0.01
```

- os.system(np.float(reg_rate))
- mlflow.log_metric('reg_rate', np.float(reg_rate))
- print(np.float(reg_rate))

- os._exit()
- mlflow.end_run()
- print(flush=True)

Correct Answer:

Answer Area

```
import numpy as np
```

- import os
- import mlflow
- import random

- os.system(f"random()")
- mlflow.start_run()
- print(random())

```
reg_rate = 0.01
```

- os.system(np.float(reg_rate))
- mlflow.log_metric('reg_rate', np.float(reg_rate))
- print(np.float(reg_rate))

- os._exit()
- mlflow.end_run()
- print(flush=True)

Question #169*Topic 3*

HOTSPOT

-

You perform hyperparameter tuning with Azure Machine Learning.

You create the following Python code:

```
command_job_for_sweep = command_job(
    learning_rate=Normal(mu=10, sigma=3),
    keep_probability=Uniform(min_value=0.05, max_value=0.1),
)
```

For each of the following statements, select Yes if the statement is true. Otherwise, select No.

Answer Area

Statements	Yes	No
The code defines a search space by using the `learning_rate` and `keep_probability` parameters.	○	○
The logarithm of the `learning_rate` parameter has a normal distribution.	○	○
The `keep_probability` parameter has a uniform distribution with a minimum value of 0.05 and a maximum value of 0.1.	○	○

Correct Answer:

Answer Area

Statements	Yes	No
The code defines a search space by using the `learning_rate` and `keep_probability` parameters.	$\boxed{\circ}$	○
The logarithm of the `learning_rate` parameter has a normal distribution.	$\boxed{\circ}$	○
The `keep_probability` parameter has a uniform distribution with a minimum value of 0.05 and a maximum value of 0.1.	$\boxed{\circ}$	○

Question #170*Topic 3*

HOTSPOT

-

You download a .csv file from a notebook in an Azure Machine Learning workspace to a data/sample.csv folder on a compute instance. The file contains 10,000 records.

You must generate the summary statistics for the data in the file. The statistics must include the following for each numerical column:

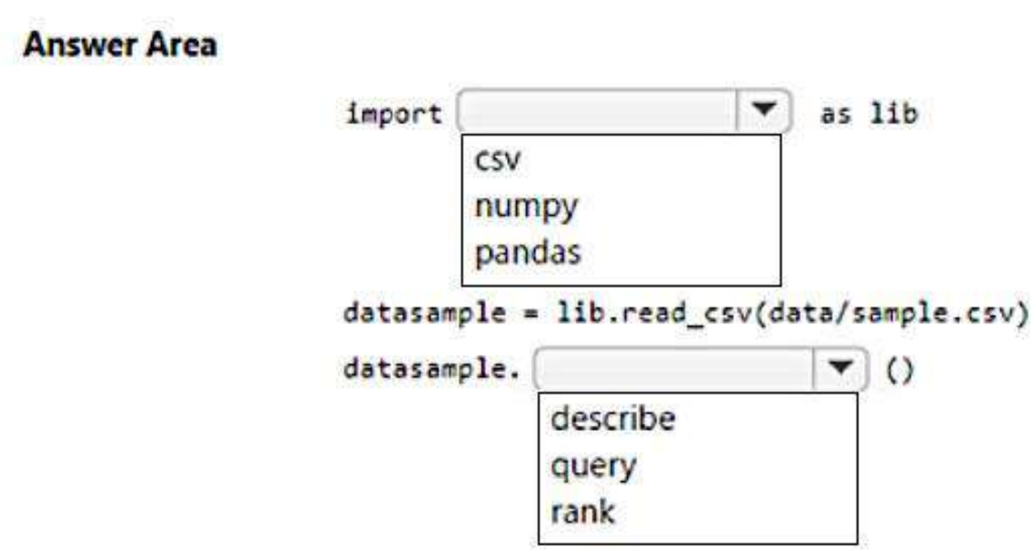
• number of non-empty values

• average value
• standard deviation
• minimum and maximum values
• 25th, 50th, and 75th percentiles

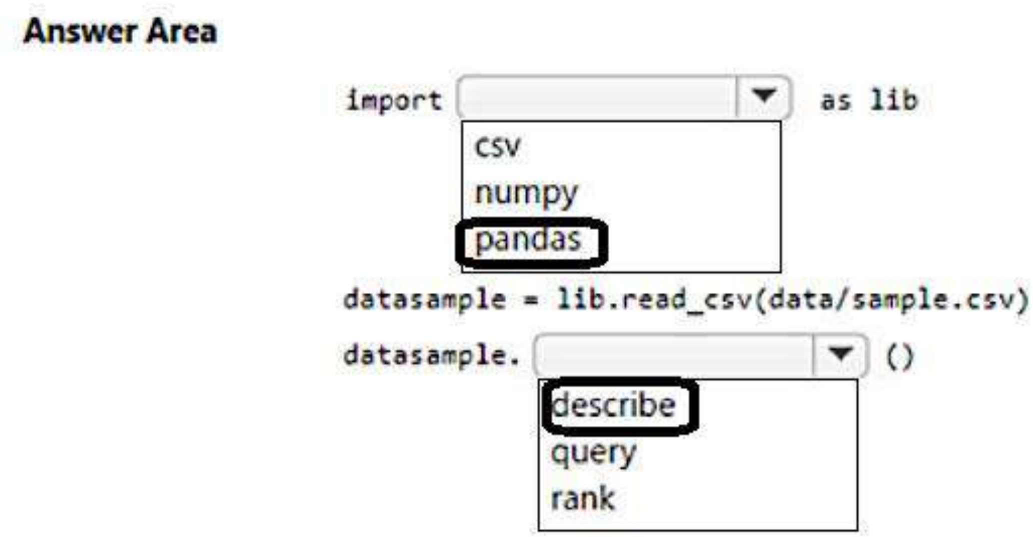
You need to complete the Python code that will generate the summary statistics.

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

NOTE: Each correct selection is worth one point.



Correct Answer:



Question #171*Topic 3*
DRAG DROP
-

You are designing an Azure Machine Learning solution by using the Python SDK v2.

You must train and deploy the solution by using a compute target. The compute target must meet the following requirements:

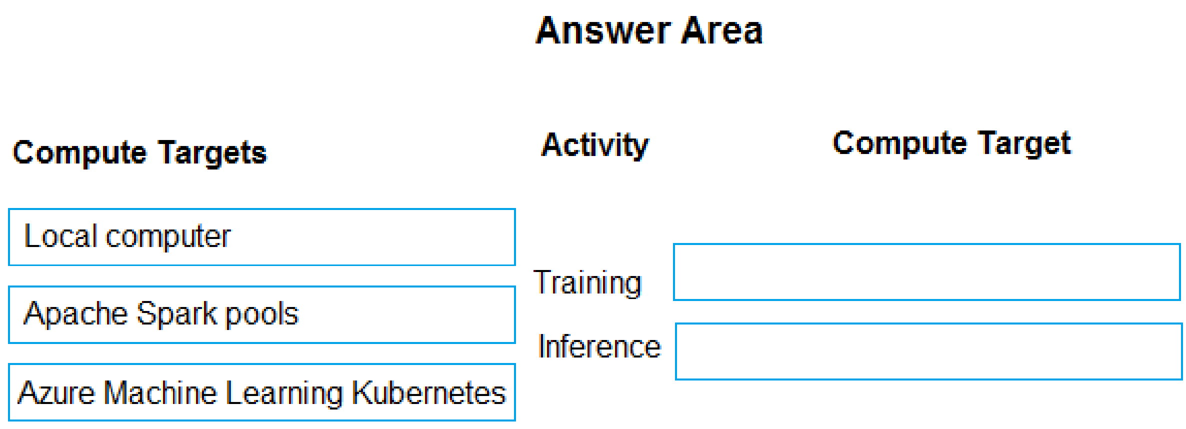
- Enable the use of on-premises compute resources.
- Support autoscaling.

You need to configure a compute target for training and inference.

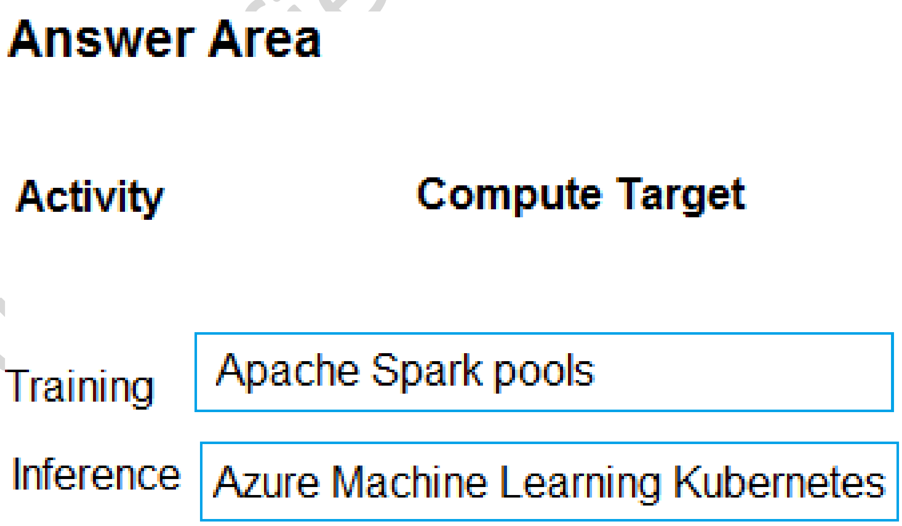
Which compute targets should you configure?

To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.



Correct Answer:



Question #172*Topic 3*

HOTSPOT

-

You create an Azure Machine Learning workspace.

You are developing a Python SDK v2 noteboot to perform custom model training in the workspace.

The notebook code imports all required packages.

You need to complete the Python SDK v2 code to include a training script, environment, and compute information.

How should you complete the code? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Answer Area

var1 = [▼] .from_config(credential=DefaultAzureCredential())

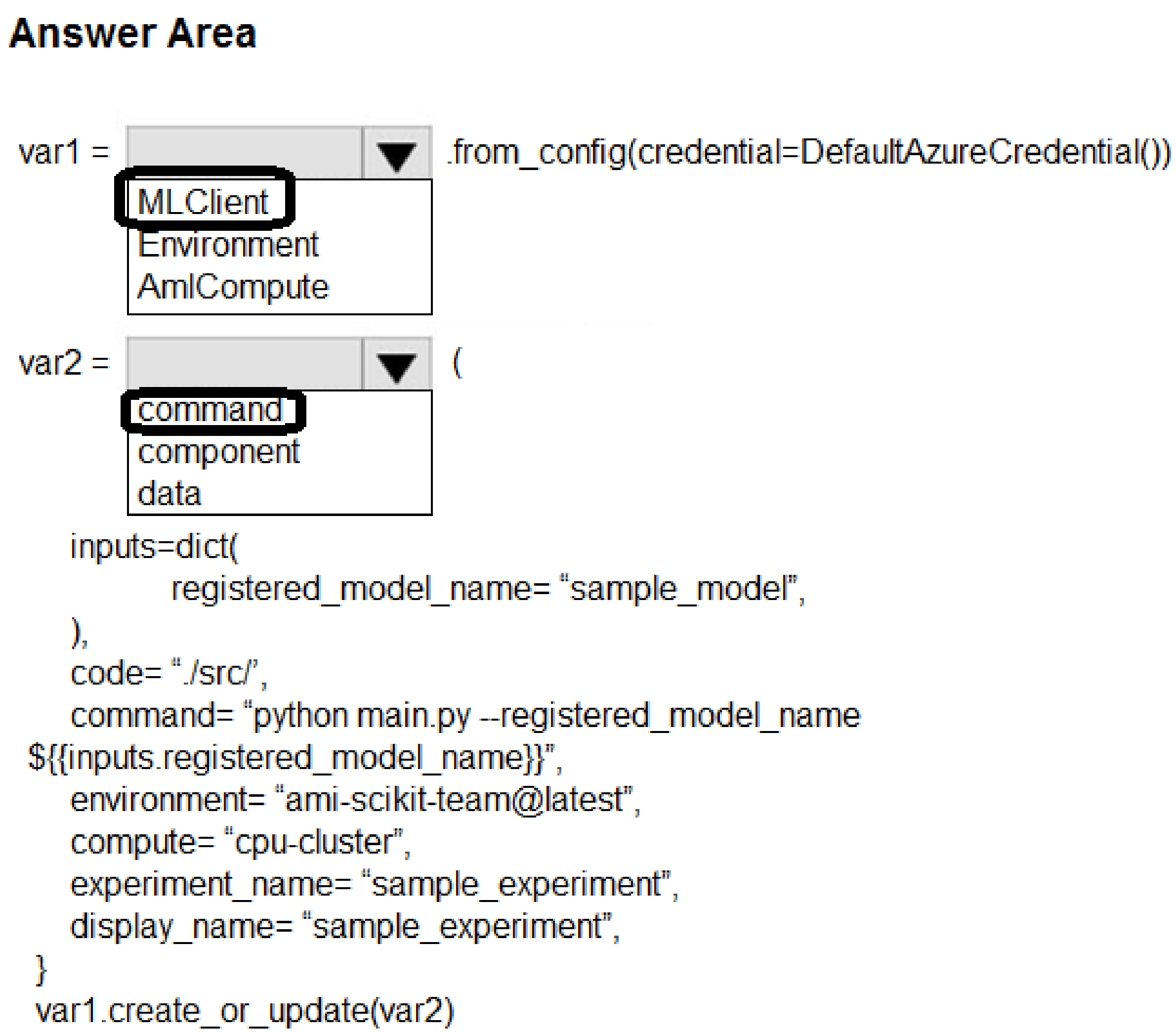
- MLClient
- Environment
- AmlCompute

var2 = [▼] (

- command
- component
- data

```
    inputs=dict(
            registered_model_name= "sample_model",
    ),
    code= "./src/",
    command= "python main.py --registered_model_name
${{inputs.registered_model_name}}",
    environment= "ami-scikit-team@latest",
    compute= "cpu-cluster",
    experiment_name= "sample_experiment",
    display_name= "sample_experiment",
 }
 var1.create_or_update(var2)
```

Correct Answer:



Question #173*Topic 3*

You have an Azure Machine Learning workspace.

You plan to tune a model hyperparameter when you train the model.

You need to define a search space that returns a normally distributed value.

Which parameter should you use?

- A. QUniform
- B. LogUniform
- C. Uniform
- D. LogNormal

Correct Answer: *D*

Question #174*Topic 3*

You have an Azure Machine Learning workspace named WS1.

You plan to use the Responsible AI dashboard to assess MLflow models that you will register in WS1.

You need to identify the library you should use to register the MLflow models.

Which library should you use?

- A. PyTorch
- B. mlpy
- C. TensorFlow
- D. scikit-learn

Correct Answer: *A*

Question #175*Topic 3*

You have an Azure Machine Learning workspace.

You plan to use the workspace to set up automated machine learning training for an image classification model.

You need to choose the primary metric to optimize the model training.

Which primary metric should you choose?

- A. r2_score
- B. mean_absolute_error
- C. accuracy
- D. root_mean_squared_log_error

Correct Answer: *C*

Question #176*Topic 3*

You have an Azure Machine Learning workspace.

You plan to use the workspace to set up automated machine learning training for an image classification model.

You need to choose the primary metric to optimize the model training.

Which primary metric should you choose?

- A. explained_variance
- B. mean_absolute_error
- C. accuracy
- D. median_absolute_error

Correct Answer: *C*

Question #: 177
Topic #: 3

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 have an Azure Machine Learning workspace.

You plan to tune model hyperparameters by using a sweep job.

You need to find a sampling method that supports early termination of low-performance jobs and continuous hyperparameters.

Solution: Use the random sampling method over the hyperparameter space.

Does the solution meet the goal?

- A. Yes
- B. No

Correct Answer: *A*

Question #: 178
Topic #: 3

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 have an Azure Machine Learning workspace.

You plan to tune model hyperparameters by using a sweep job.

You need to find a sampling method that supports early termination of low-performance jobs and continuous hyperparameters.

Solution: Use the Bayesian sampling method over the hyperparameter space.

Does the solution meet the goal?

- A. Yes
- B. No

Correct Answer: *B*

Question #: 179
Topic #: 3

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 have an Azure Machine Learning workspace.

You plan to tune model hyperparameters by using a sweep job.

You need to find a sampling method that supports early termination of low-performance jobs and continuous hyperparameters.

Solution: Use the grid sampling method over the hyperparameter space.

Does the solution meet the goal?

- A. Yes
- B. No

Correct Answer: *B*

Question #: 180
Topic #: 3

You have an Azure Machine Learning workspace.

You plan to use the workspace to set up automated machine learning training for an image classification model.

You need to choose the primary metric to optimize the model training.

Which primary metric should you choose?

- A. r2_score
- B. mean_absolute_error
- C. iou

- D. median_absolute_error

Correct Answer: *C*

Question #: 181
Topic #: 3

You have an Azure Machine Learning workspace named WS1.

You plan to use Azure Machine Learning SDK v2 to register a model as an asset in WS1 from an artifact generated by an MLflow run. The artifact resides in a named output of a job used for the model training.

You need to identify the syntax of the path to reference the model when you register it.

Which syntax should you use?

- A. t//model/
- B. azureml://registries
- C. mlflow-model/
- D. azureml://jobs/

Correct Answer: *A*

Question #: 182
Topic #: 3

You manage an Azure Machine Learning workspace.

You experiment with an MLflow model that trains interactively by using a notebook in the workspace.

You need to log dictionary type artifacts of the experiments in Azure Machine Learning by using MLflow.

Which syntax should you use?

- A. mlflow.log_input(my_dict)
- B. mlflow.log_metric("my_metric", my_dict)
- C. mlflow.log_metrics(my_dict)
- D. mlflow.log_text("my_metric", my_dict)

Correct Answer: *D*

Question #: 183
Topic #: 3

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 manage an Azure Machine Learning workspace. The Python script named script.py reads an argument named training_data. The training_data argument specifies the path to the training data in a file named dataset1.csv.

You plan to run the script.py Python script as a command job that trains a machine learning model.

You need to provide the command to pass the path for the dataset as a parameter value when you submit the script as a training job.

Solution: python train.py --training_data training_data

Does the solution meet the goal?

- A. Yes
- B. No

Correct Answer: *B*

Question #: 184
Topic #: 3

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 manage an Azure Machine Learning workspace. The Python script named script.py reads an argument named training_data. The training_data argument specifies the path to the training data in a file named dataset1.csv.

You plan to run the script.py Python script as a command job that trains a machine learning model.

You need to provide the command to pass the path for the dataset as a parameter value when you submit the script as a training job.

Solution: python script.py dataset1.csv

Does the solution meet the goal?

- A. Yes
- B. No

Correct Answer: *B*

Question #: 185
Topic #: 3

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 manage an Azure Machine Learning workspace. The Python script named script.py reads an argument named training_data. The training_data argument specifies the path to the training data in a file named dataset1.csv.

You plan to run the script.py Python script as a command job that trains a machine learning model.

You need to provide the command to pass the path for the dataset as a parameter value when you submit the script as a training job.

Solution: python script.py --training_data ${{inputs.training_data}}

Does the solution meet the goal?

- A. Yes
- B. No

Correct Answer: *A*

Question #: 186
Topic #: 3

HOTSPOT
-

You manage an Azure Machine Learning workspace.

You plan to train a natural language processing (NLP) model that will assign labels for designated tokens in unstructured text.

You need to configure the NLP task by using automated machine learning.

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

NOTE: Each correct selection is worth one point.

Answer Area

NLP task configuration

Configuration setting	Value
Task	Named Entity Recognition Multi-class text classification Multi-label text classification
File format	CSV JSON CoNLL

Correct Answer:

Answer Area

NLP task configuration

Configuration setting	Value
Task	Named Entity Recognition Multi-class text classification Multi-label text classification
File format	CSV JSON CoNLL

Question #: 187
Topic #: 3

HOTSPOT
-

You manage an Azure Machine Learning workspace. The development environment is configured with a Serverless Spark compute in Azure Machine Learning Notebooks.

You perform interactive data wrangling to clean up the Titanic dataset and store it as a new dataset. (Line numbers are used for reference only.)

```
01 import pyspark.pandas as pd
02 from pyspark.ml.feature import Imputer
03
04 df = pd.read_csv("abfss://<FILE_SYSTEM_NAME>@<STORAGE_ACCOUNT_NAME>.dfs.core.windows.net/data/titanic.csv",
index_col="PassengerId")
05
06 imputer = Imputer(inputCols=["Age"], outputCol="Age").setStrategy("mean")
07 df.fillna(value={"Cabin": "None"}, inplace=False)
08 df.dropna(inplace=True)
09
10 df.to_csv( "abfss://<FILE_SYSTEM_NAME>@<STORAGE_ACCOUNT_NAME>.dfs.core.windows.net/data/wrangled",
index_col="PassengerId")
```

For each of the following statements, select Yes if the statement is true. Otherwise, select No.

NOTE: Each correct selection is worth one point.

Answer Area

Interactive data wrangling with Apache Spark

Statement	Yes	No
All missing values in the Age column are filled with the mean value.	○	○
Cabin column is filled with None if data is missing.	○	○
Data is imported from Azure Data Lake Storage Gen2.	○	○
User identity passthrough is enabled to access data.	○	○

Correct Answer:

Answer Area

Interactive data wrangling with Apache Spark

Statement	Yes	No
All missing values in the Age column are filled with the mean value.	$\boxed{\circ}$	○
Cabin column is filled with None if data is missing.	$\boxed{\circ}$	○
Data is imported from Azure Data Lake Storage Gen2.	$\boxed{\circ}$	○
User identity passthrough is enabled to access data.	$\boxed{\circ}$	○

Question #: 190
Topic #: 3

You manage an Azure Machine Learning workspace.

You choose the uri_folder data type as an output of a pipeline component.

You need to define the data access mode that is supported by your configuration.

Which mode should you define?

- A. eval_upload
- B. rw_mount
- C. download

- D. ro_mount

Correct Answer: *B*

Question #: 191
Topic #: 3

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 manage an Azure Machine Learning workspace. The Python script named script.py reads an argument named training_data. The training_data argument specifies the path to the training data in a file named dataset1.csv.

You plan to run the script.py Python script as a command job that trains a machine learning model.

You need to provide the command to pass the path for the dataset as a parameter value when you submit the script as a training job.

Solution: python script.py --training_data dataset1.csv

Does the solution meet the goal?

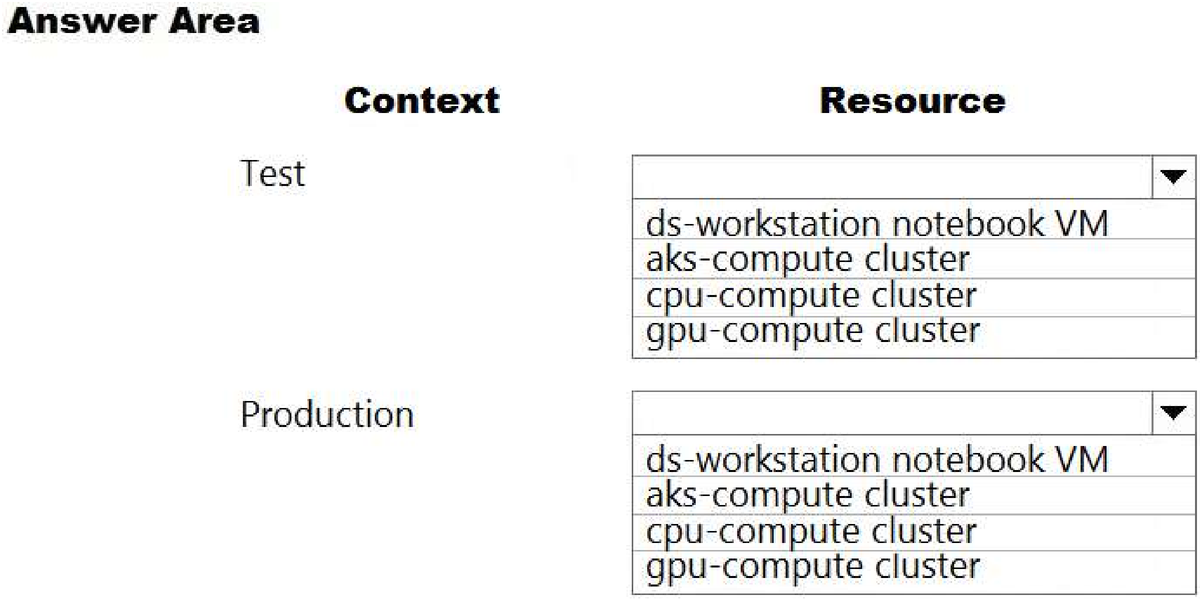
- A. Yes
- B. No

Correct Answer: *A*

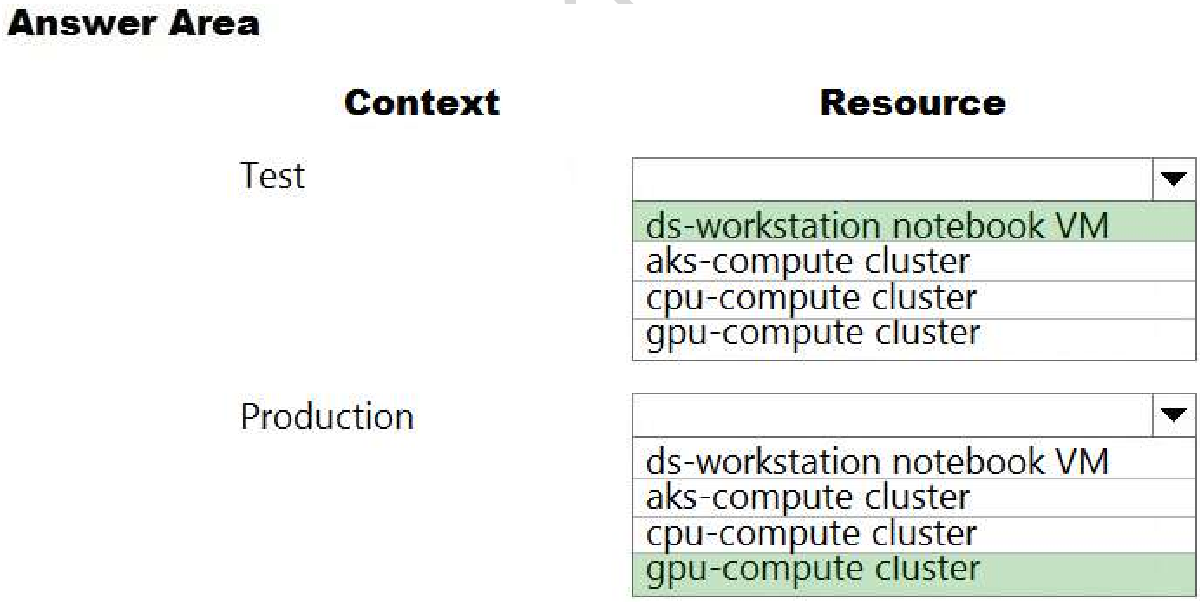
Topic 4 - Question Set 4
Question #1*Topic 4*

HOTSPOT -
You are a lead data scientist for a project that tracks the health and migration of birds. You create a multi-image classification deep learning model that uses a set of labeled bird photos collected by experts. You plan to use the model to develop a cross-platform mobile app that predicts the species of bird captured by app users.
You must test and deploy the trained model as a web service. The deployed model must meet the following requirements:
⇨ An authenticated connection must not be required for testing.
⇨ The deployed model must perform with low latency during inferencing.
⇨ The REST endpoints must be scalable and should have a capacity to handle large number of requests when multiple end users are using the mobile application.

You need to verify that the web service returns predictions in the expected JSON format when a valid REST request is submitted.
Which compute resources should you use? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.
Hot Area:



Correct Answer:



Box 1: ds-workstation notebook VM
An authenticated connection must not be required for testing.
On a Microsoft Azure virtual machine (VM), including a Data Science Virtual Machine (DSVM), you create local user accounts while provisioning the VM. Users then authenticate to the VM by using these credentials.

Box 2: gpu-compute cluster -

Image classification is well suited for GPU compute clusters
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/data-science-virtual-machine/dsvm-common-identity https://docs.microsoft.com/en-us/azure/architecture/reference-architectures/ai/training-deep-learning
Question #2*Topic 4*

You create a deep learning model for image recognition on Azure Machine Learning service using GPU-based training.
You must deploy the model to a context that allows for real-time GPU-based inferencing.
You need to configure compute resources for model inferencing.
Which compute type should you use?

- A. Azure Container Instance
- B. Azure Kubernetes Service
- C. Field Programmable Gate Array
- D. Machine Learning Compute

Correct Answer: *B*
You can use Azure Machine Learning to deploy a GPU-enabled model as a web service. Deploying a model on Azure Kubernetes Service (AKS) is one option.
The AKS cluster provides a GPU resource that is used by the model for inference.
Inference, or model scoring, is the phase where the deployed model is used to make predictions.
Using GPUs instead of CPUs offers performance advantages on highly parallelizable computation.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-deploy-inferencing-gpus
Question #3*Topic 4*

You create a batch inference pipeline by using the Azure ML SDK. You run the pipeline by using the following code: from azureml.pipeline.core import Pipeline from azureml.core.experiment import Experiment pipeline = Pipeline(workspace=ws, steps=[parallelrun_step]) pipeline_run = Experiment(ws, 'batch_pipeline').submit(pipeline)
You need to monitor the progress of the pipeline execution.
What are two possible ways to achieve this goal? Each correct answer presents a complete solution.
NOTE: Each correct selection is worth one point.

- A. Run the following code in a notebook:
```
from azureml.contrib.interpret.explanation.explanation_client import ExplanationClient
client = ExplanationClient.from_run(pipeline_run)
explanation = client.download_model_explanation()
explanation = client.download_model_explanation(top_k=4)
global_importance_values = explanation.get_ranked_global_values()
global_importance_names = explanation.get_ranked_global_names()
print('global importance values: {}'.format(global_importance_values))
print('global importance names: {}'.format(global_importance_names))
```
- B. Use the Inference Clusters tab in Machine Learning Studio.
- C. Use the Activity log in the Azure portal for the Machine Learning workspace.
- D. Run the following code in a notebook:
```
from azureml.widgets import RunDetails
RunDetails(pipeline_run).show()
```

- E. Run the following code and monitor the console output from the PipelineRun object:
```
pipeline_run.wait_for_completion(show_output=True)
```

Correct Answer: *DE*
A batch inference job can take a long time to finish. This example monitors progress by using a Jupyter widget. You can also manage the job's progress by using:
⇨ Azure Machine Learning Studio.
⇨ Console output from the PipelineRun object.
from azureml.widgets import RunDetails
RunDetails(pipeline_run).show()
pipeline_run.wait_for_completion(show_output=True)
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-use-parallel-run-step#monitor-the-parallel-run-job
Question #4*Topic 4*

You train and register a model in your Azure Machine Learning workspace.
You must publish a pipeline that enables client applications to use the model for batch inferencing.
You must use a pipeline with a single ParallelRunStep step that runs a Python inferencing script to get predictions from the input data.
You need to create the inferencing script for the ParallelRunStep pipeline step.
Which two functions should you include? Each correct answer presents part of the solution.
NOTE: Each correct selection is worth one point.

- A. run(mini_batch)
- B. main()
- C. batch()
- D. init()
- E. score(mini_batch)

Correct Answer: *AD*
Reference:
https://github.com/Azure/MachineLearningNotebooks/tree/master/how-to-use-azureml/machine-learning-pipelines/parallel-run

Question #5*Topic 4*

You deploy a model as an Azure Machine Learning real-time web service using the following code.
```
# ws, model, interference_config, and deployment_config defined previously
service = Model.deploy(ws, 'classification-service', [model], inference_config, deployment_config)
service.wait_for_deployment(True)
```
The deployment fails.
You need to troubleshoot the deployment failure by determining the actions that were performed during deployment and identifying the specific action that failed.
Which code segment should you run?

- A. service.get_logs()
- B. service.state
- C. service.serialize()
- D. service.update_deployment_state()

Correct Answer: *A*

You can print out detailed Docker engine log messages from the service object. You can view the log for ACI, AKS, and Local deployments. The following example demonstrates how to print the logs.

if you already have the service object handy

print(service.get_logs())

if you only know the name of the service (note there might be multiple services with the same name but different version number) print(ws.webservices['mysvc'].get_logs())

Reference:

https://docs.microsoft.com/en-us/azure/machine-learning/how-to-troubleshoot-deployment

Question #6*Topic 4*

HOTSPOT -

You deploy a model in Azure Container Instance.

You must use the Azure Machine Learning SDK to call the model API.

You need to invoke the deployed model using native SDK classes and methods.

How should you complete the command? To answer, select the appropriate options in the answer areas.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

```
from azureml.core import Workspace
```

from azureml.core.webservice import requests
from azureml.core.webservice import Webservice
from azureml.core.webservice import LocalWebservice

```
import json
ws = Workspace.from_config()
service_name = "mlmodel1-service"
service = Webservice(name=service_name, workspace=ws)
x_new = [[2,101.5,1,24,21], [1,89.7,4,41,21]]
input_json = json.dumps({"data": x_new})
```

predictions = service.run(input_json)
predictions = requests.post(service.scoring_uri, input_json)
predictions = service.deserialize(ws, input_json)

Correct Answer:

Answer Area

```
from azureml.core import Workspace
```

from azureml.core.webservice import requests
from azureml.core.webservice import Webservice
from azureml.core.webservice import LocalWebservice

```
import json
ws = Workspace.from_config()
service_name = "mlmodel1-service"
service = Webservice(name=service_name, workspace=ws)
x_new = [[2,101.5,1,24,21], [1,89.7,4,41,21]]
input_json = json.dumps({"data": x_new})
```

predictions = service.run(input_json)
predictions = requests.post(service.scoring_uri, input_json)
predictions = service.deserialize(ws, input_json)

Box 1: from azureml.core.webservice import Webservice
The following code shows how to use the SDK to update the model, environment, and entry script for a web service to Azure Container Instances: from azureml.core import Environment from azureml.core.webservice import Webservice from azureml.core.model import Model, InferenceConfig
Box 2: predictions = service.run(input_json)
Example: The following code demonstrates sending data to the service: import json test_sample = json.dumps({'data': [
[1, 2, 3, 4, 5, 6, 7, 8, 9, 10],
[10, 9, 8, 7, 6, 5, 4, 3, 2, 1]
]})
test_sample = bytes(test_sample, encoding='utf8')
prediction = service.run(input_data=test_sample)
print(prediction)
Reference:
https://docs.microsoft.com/bs-latn-ba/azure/machine-learning/how-to-deploy-azure-container-instance https://docs.microsoft.com/en-us/azure/machine-learning/how-to-troubleshoot-deployment
Question #7*Topic 4*

You create a multi-class image classification deep learning model.
You train the model by using PyTorch version 1.2.
You need to ensure that the correct version of PyTorch can be identified for the inferencing environment when the model is deployed.
What should you do?

- A. Save the model locally as a.pt file, and deploy the model as a local web service.
- B. Deploy the model on computer that is configured to use the default Azure Machine Learning conda environment.
- C. Register the model with a .pt file extension and the default version property.

- D. Register the model, specifying the model_framework and model_framework_version properties.

Correct Answer: *D*
framework_version: The PyTorch version to be used for executing training code.
Reference:
https://docs.microsoft.com/en-us/python/api/azureml-train-core/azureml.train.dnn.pytorch?view=azure-ml-py

Question #8*Topic 4*

You train a machine learning model.
You must deploy the model as a real-time inference service for testing. The service requires low CPU utilization and less than 48 MB of RAM. The compute target for the deployed service must initialize automatically while minimizing cost and administrative overhead.
Which compute target should you use?

- A. Azure Container Instance (ACI)
- B. attached Azure Databricks cluster
- C. Azure Kubernetes Service (AKS) inference cluster
- D. Azure Machine Learning compute cluster

Correct Answer: *A*
Azure Container Instances (ACI) are suitable only for small models less than 1 GB in size.
Use it for low-scale CPU-based workloads that require less than 48 GB of RAM.
Note: Microsoft recommends using single-node Azure Kubernetes Service (AKS) clusters for dev-test of larger models.
Reference:
https://docs.microsoft.com/id-id/azure/machine-learning/how-to-deploy-and-where

Question #9*Topic 4*

You register a model that you plan to use in a batch inference pipeline.
The batch inference pipeline must use a ParallelRunStep step to process files in a file dataset. The script has the ParallelRunStep step runs must process six input files each time the inferencing function is called.
You need to configure the pipeline.
Which configuration setting should you specify in the ParallelRunConfig object for the PrallelRunStep step?

- A. process_count_per_node= "6"
- B. node_count= "6"
- C. mini_batch_size= "6"
- D. error_threshold= "6"

Correct Answer: *C*
node_count is the number of nodes in the compute target used for running the ParallelRunStep.
Incorrect Answers:

A: process_count_per_node -

Number of processes executed on each node. (optional, default value is number of cores on node.)

C: mini_batch_size -
For FileDataset input, this field is the number of files user script can process in one run() call. For TabularDataset input, this field is the approximate size of data the user script can process in one run() call. Example values are 1024, 1024KB, 10MB, and 1GB.

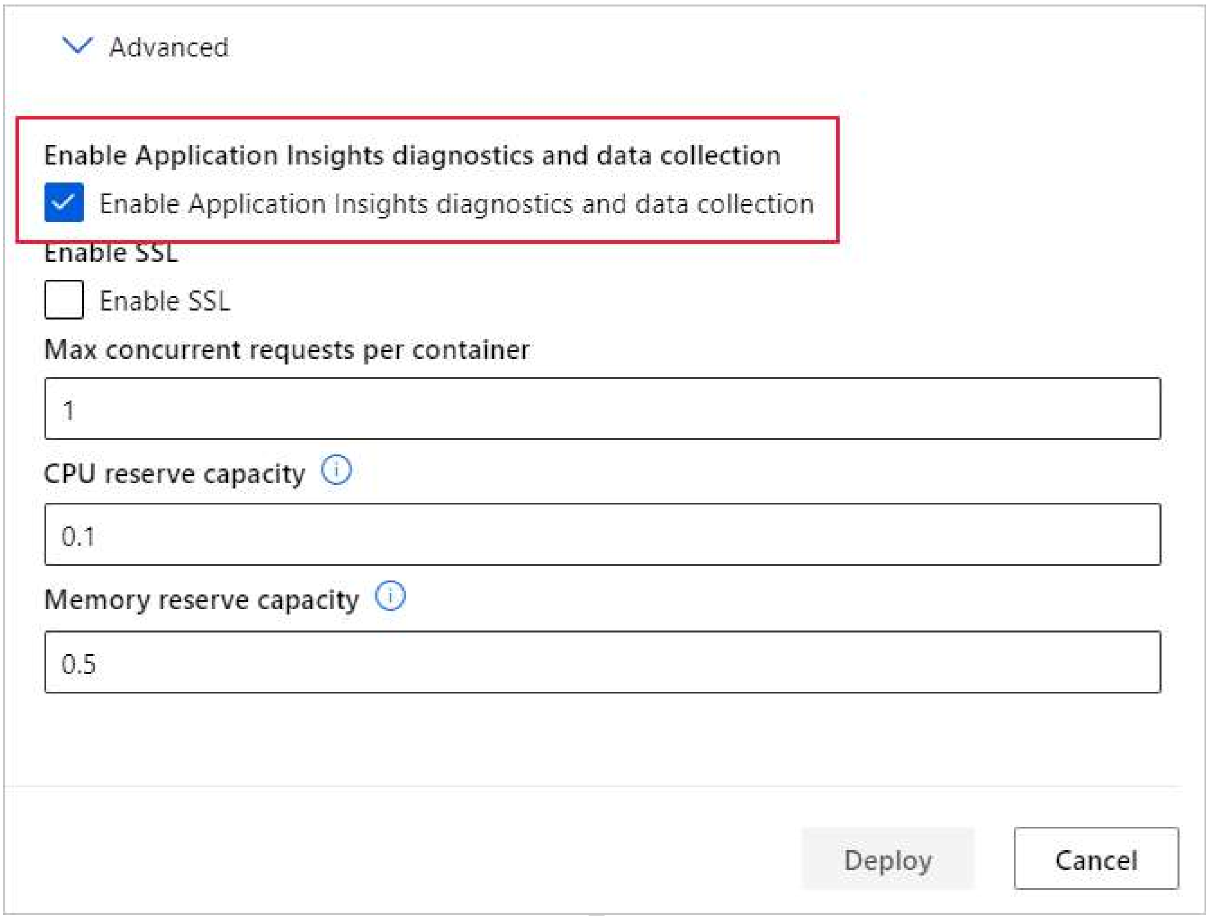
D: error_threshold -
The number of record failures for TabularDataset and file failures for FileDataset that should be ignored during processing. If the error count goes above this value, then the job will be aborted.
Reference:
https://docs.microsoft.com/en-us/python/api/azureml-contrib-pipeline-steps/azureml.contrib.pipeline.steps.parallelrunconfig?view=azure-ml-py

Question #10*Topic 4*

You deploy a real-time inference service for a trained model.
The deployed model supports a business-critical application, and it is important to be able to monitor the data submitted to the web service and the predictions the data generates.
You need to implement a monitoring solution for the deployed model using minimal administrative effort.
What should you do?

- A. View the explanations for the registered model in Azure ML studio.
- B. Enable Azure Application Insights for the service endpoint and view logged data in the Azure portal.
- C. View the log files generated by the experiment used to train the model.
- D. Create an ML Flow tracking URI that references the endpoint, and view the data logged by ML Flow.

Correct Answer: *B*
Configure logging with Azure Machine Learning studio
You can also enable Azure Application Insights from Azure Machine Learning studio. When you're ready to deploy your model as a web service, use the following steps to enable Application Insights:
1. Sign in to the studio at https://ml.azure.com.
2. Go to Models and select the model you want to deploy.
3. Select +Deploy.
4. Populate the Deploy model form.
5. Expand the Advanced menu.
6. Select Enable Application Insights diagnostics and data collection.



Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-enable-app-insights

Question #11*Topic 4*

HOTSPOT -
You use Azure Machine Learning to train and register a model.
You must deploy the model into production as a real-time web service to an inference cluster named service-compute that the IT department has created in the
Azure Machine Learning workspace.
Client applications consuming the deployed web service must be authenticated based on their Azure Active Directory service principal.
You need to write a script that uses the Azure Machine Learning SDK to deploy the model. The necessary modules have been imported.
How should you complete the code? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

```
# Assume the necessary modules have been imported
deploy_target = [ ▼ ] .(ws, "service-compute")
```

- AksCompute
- AmlCompute
- RemoteCompute
- BatchCompute

```
deployment_config = [ ▼ ] .deploy_configuration(cpu_cores=1, memory_gb=1,
```

- AksWebservice
- AciWebservice
- LocalWebService

```
[ ▼ ] )
```

- token_auth_enabled=True
- token_auth_enabled=False
- auth_enabled=True
- auth_enabled=False

```
service = Model.deploy(ws, "ml-service",
  [model], inference_config, deployment_config, deploy_target)
service.wait_for_deployment(show_output = True)
```

Correct Answer:

Answer Area

```
# Assume the necessary modules have been imported
deploy_target = [ ▼ ] .(ws, "service-compute")
```

- AksCompute
- AmlCompute
- RemoteCompute
- BatchCompute

```
deployment_config = [ ▼ ] .deploy_configuration(cpu_cores=1, memory_gb=1,
```

- AksWebservice
- AciWebservice
- LocalWebService

```
[ ▼ ] )
```

- token_auth_enabled=True
- token_auth_enabled=False
- auth_enabled=True
- auth_enabled=False

```
service = Model.deploy(ws, "ml-service",
  [model], inference_config, deployment_config, deploy_target)
service.wait_for_deployment(show_output = True)
```

Box 1: AksCompute -

Example:

aks_target = AksCompute(ws,"myaks")

If deploying to a cluster configured for dev/test, ensure that it was created with enough

cores and memory to handle this deployment configuration. Note that memory is also used by

things such as dependencies and AML components.

deployment_config = AksWebservice.deploy_configuration(cpu_cores = 1, memory_gb = 1) service = Model.deploy(ws, "myservice", [model], inference_config, deployment_config, aks_target)

Box 2: AksWebservice -

Box 3: token_auth_enabled=Yes -

Whether or not token auth is enabled for the Webservice.
Note: A Service principal defined in Azure Active Directory (Azure AD) can act as a principal on which authentication and authorization policies can be enforced in
Azure Databricks.
The Azure Active Directory Authentication Library (ADAL) can be used to programmatically get an Azure AD access token for a user.
Incorrect Answers:
auth_enabled (bool): Whether or not to enable key auth for this Webservice. Defaults to True.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-deploy-azure-kubernetes-service
https://docs.microsoft.com/en-us/azure/databricks/dev-tools/api/latest/aad/service-prin-aad-token
Question #12*Topic 4*

An organization creates and deploys a multi-class image classification deep learning model that uses a set of labeled photographs.
The software engineering team reports there is a heavy inferencing load for the prediction web services during the summer. The production web service for the model fails to meet demand despite having a fully-utilized compute cluster where the web service is deployed.
You need to improve performance of the image classification web service with minimal downtime and minimal administrative effort.
What should you advise the IT Operations team to do?

- A. Create a new compute cluster by using larger VM sizes for the nodes, redeploy the web service to that cluster, and update the DNS registration for the service endpoint to point to the new cluster.
- B. Increase the node count of the compute cluster where the web service is deployed.
- C. Increase the minimum node count of the compute cluster where the web service is deployed.
- D. Increase the VM size of nodes in the compute cluster where the web service is deployed.

Correct Answer: *B*
The Azure Machine Learning SDK does not provide support scaling an AKS cluster. To scale the nodes in the cluster, use the UI for your AKS cluster in the Azure
Machine Learning studio. You can only change the node count, not the VM size of the cluster.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-create-attach-kubernetes

Question #13*Topic 4*

You use Azure Machine Learning designer to create a real-time service endpoint. You have a single Azure Machine Learning service compute resource.
You train the model and prepare the real-time pipeline for deployment.
You need to publish the inference pipeline as a web service.
Which compute type should you use?

- A. a new Machine Learning Compute resource
- B. Azure Kubernetes Services
- C. HDInsight
- D. the existing Machine Learning Compute resource
- E. Azure Databricks

Correct Answer: *B*
Azure Kubernetes Service (AKS) can be used real-time inference.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/concept-compute-target

Question #14*Topic 4*

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 train and register a machine learning model.
You plan to deploy the model as a real-time web service. Applications must use key-based authentication to use the model.
You need to deploy the web service.
Solution:
Create an AciWebservice instance.
Set the value of the ssl_enabled property to True.
Deploy the model to the service.
Does the solution meet the goal?

- A. Yes
- B. No

Correct Answer: *B*
Instead use only auth_enabled = TRUE
Note: Key-based authentication.
Web services deployed on AKS have key-based auth enabled by default. ACI-deployed services have key-based auth disabled by default, but you can enable it by setting auth_enabled = TRUE when creating the ACI web service. The following is an example of creating an ACI deployment configuration with key-based auth enabled. deployment_config <-
aci_webservice_deployment_config(cpu_cores = 1, memory_gb = 1, auth_enabled = TRUE)
Reference:
https://azure.github.io/azureml-sdk-for-r/articles/deploying-models.html

Question #15*Topic 4*

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 train and register a machine learning model.
You plan to deploy the model as a real-time web service. Applications must use key-based authentication to use the model.
You need to deploy the web service.
Solution:
Create an AciWebservice instance.
Set the value of the auth_enabled property to True.

Deploy the model to the service.
Does the solution meet the goal?

- A. Yes
- B. No

Correct Answer: *A*
Key-based authentication.
Web services deployed on AKS have key-based auth enabled by default. ACI-deployed services have key-based auth disabled by default, but you can enable it by setting auth_enabled = TRUE when creating the ACI web service. The following is an example of creating an ACI deployment configuration with key-based auth enabled. deployment_config <- aci_webservice_deployment_config(cpu_cores = 1, memory_gb = 1, auth_enabled = TRUE)
Reference:
https://azure.github.io/azureml-sdk-for-r/articles/deploying-models.html

Question #16*Topic 4*

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 train and register a machine learning model.
You plan to deploy the model as a real-time web service. Applications must use key-based authentication to use the model.
You need to deploy the web service.
Solution:
Create an AciWebservice instance.
Set the value of the auth_enabled property to False.
Set the value of the token_auth_enabled property to True.
Deploy the model to the service.
Does the solution meet the goal?

- A. Yes
- B. No

Correct Answer: *B*
Instead use only auth_enabled = TRUE
Note: Key-based authentication.
Web services deployed on AKS have key-based auth enabled by default. ACI-deployed services have key-based auth disabled by default, but you can enable it by setting auth_enabled = TRUE when creating the ACI web service. The following is an example of creating an ACI deployment configuration with key-based auth enabled. deployment_config <- aci_webservice_deployment_config(cpu_cores = 1, memory_gb = 1, auth_enabled = TRUE)
Reference:
https://azure.github.io/azureml-sdk-for-r/articles/deploying-models.html

Question #17*Topic 4*

You use the following Python code in a notebook to deploy a model as a web service: from azureml.core.webservice import AciWebservice from azureml.core.model import InferenceConfig inference_config = InferenceConfig(runtime='python', source_directory='model_files', entry_script='score.py', conda_file='env.yml') deployment_config = AciWebservice.deploy_configuration(cpu_cores=1, memory_gb=1) service = Model.deploy(ws, 'my-service', [model], inference_config, deployment_config) service.wait_for_deployment(True)
The deployment fails.
You need to use the Python SDK in the notebook to determine the events that occurred during service deployment an initialization.
Which code segment should you use?

- A. service.state
- B. service.get_logs()
- C. service.serialize()
- D. service.environment

Correct Answer: *B*
The first step in debugging errors is to get your deployment logs. In Python: service.get_logs()
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-troubleshoot-deployment
Question #18*Topic 4*

You use the Azure Machine Learning Python SDK to define a pipeline that consists of multiple steps. When you run the pipeline, you observe that some steps do not run. The cached output from a previous run is used instead.
You need to ensure that every step in the pipeline is run, even if the parameters and contents of the source directory have not changed since the previous run.
What are two possible ways to achieve this goal? Each correct answer presents a complete solution.
NOTE: Each correct selection is worth one point.

- A. Use a PipelineData object that references a datastore other than the default datastore.
- B. Set the regenerate_outputs property of the pipeline to True.
- C. Set the allow_reuse property of each step in the pipeline to False.
- D. Restart the compute cluster where the pipeline experiment is configured to run.
- E. Set the outputs property of each step in the pipeline to True.

Correct Answer: *BC*
B: If regenerate_outputs is set to True, a new submit will always force generation of all step outputs, and disallow data reuse for any step of this run. Once this run is complete, however, subsequent runs may reuse the results of this run.
C: Keep the following in mind when working with pipeline steps, input/output data, and step reuse.
⇨ If data used in a step is in a datastore and allow_reuse is True, then changes to the data change won't be detected. If the data is uploaded as part of the snapshot (under the step's source_directory), though this is not recommended, then the hash will change and will trigger a rerun.
Reference:
https://docs.microsoft.com/en-us/python/api/azureml-pipeline-core/azureml.pipeline.core.pipelinestep
https://github.com/Azure/MachineLearningNotebooks/blob/master/how-to-use-azureml/machine-learning-pipelines/intro-to-pipelines/aml-pipelines-getting- started.ipynb

Question #19*Topic 4*

You train a model and register it in your Azure Machine Learning workspace. You are ready to deploy the model as a real-time web service.
You deploy the model to an Azure Kubernetes Service (AKS) inference cluster, but the deployment fails because an error occurs when the service runs the entry script that is associated with the model deployment.
You need to debug the error by iteratively modifying the code and reloading the service, without requiring a re-deployment of the service for each code update.
What should you do?

- A. Modify the AKS service deployment configuration to enable application insights and re-deploy to AKS.
- B. Create an Azure Container Instances (ACI) web service deployment configuration and deploy the model on ACI.
- C. Add a breakpoint to the first line of the entry script and redeploy the service to AKS.
- D. Create a local web service deployment configuration and deploy the model to a local Docker container.
- E. Register a new version of the model and update the entry script to load the new version of the model from its registered path.

Correct Answer: *D*
How to work around or solve common Docker deployment errors with Azure Container Instances (ACI) and Azure Kubernetes Service (AKS) using Azure Machine
Learning.
The recommended and the most up to date approach for model deployment is via the Model.deploy() API using an Environment object as an input parameter. In this case our service will create a base docker image for you during deployment stage and mount the required models all in one call. The basic deployment tasks are:
1. Register the model in the workspace model registry.
2. Define Inference Configuration:
a) Create an Environment object based on the dependencies you specify in the environment yaml file or use one of our procured environments. b) Create an inference configuration (InferenceConfig object) based on the environment and the scoring script.
3. Deploy the model to Azure Container Instance (ACI) service or to Azure Kubernetes Service (AKS).

Question #20*Topic 4*
You use Azure Machine Learning designer to create a training pipeline for a regression model.
You need to prepare the pipeline for deployment as an endpoint that generates predictions asynchronously for a dataset of input data values.
What should you do?

- A. Clone the training pipeline.
- B. Create a batch inference pipeline from the training pipeline.
- C. Create a real-time inference pipeline from the training pipeline.
- D. Replace the dataset in the training pipeline with an Enter Data Manually module.

Correct Answer: *B*
You must first convert the training pipeline into a real-time inference pipeline. This process removes training modules and adds web service inputs and outputs to handle requests.
Incorrect Answers:
A: Use the Enter Data Manually module to create a small dataset by typing values.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/tutorial-designer-automobile-price-deploy https://docs.microsoft.com/en-us/azure/machine-learning/algorithm-module-reference/enter-data-manually

Question #21*Topic 4*

You retrain an existing model.
You need to register the new version of a model while keeping the current version of the model in the registry.
What should you do?

- A. Register a model with a different name from the existing model and a custom property named version with the value 2.
- B. Register the model with the same name as the existing model.
- C. Save the new model in the default datastore with the same name as the existing model. Do not register the new model.
- D. Delete the existing model and register the new one with the same name.

Correct Answer: *B*
Model version: A version of a registered model. When a new model is added to the Model Registry, it is added as Version 1. Each model registered to the same model name increments the version number.
Reference:
https://docs.microsoft.com/en-us/azure/databricks/applications/mlflow/model-registry
Question #22*Topic 4*

You use the Azure Machine Learning SDK to run a training experiment that trains a classification model and calculates its accuracy metric.
The model will be retrained each month as new data is available.
You must register the model for use in a batch inference pipeline.
You need to register the model and ensure that the models created by subsequent retraining experiments are registered only if their accuracy is higher than the currently registered model.
What are two possible ways to achieve this goal? Each correct answer presents a complete solution.
NOTE: Each correct selection is worth one point.

- A. Specify a different name for the model each time you register it.
- B. Register the model with the same name each time regardless of accuracy, and always use the latest version of the model in the batch inferencing pipeline.
- C. Specify the model framework version when registering the model, and only register subsequent models if this value is higher.
- D. Specify a property named accuracy with the accuracy metric as a value when registering the model, and only register subsequent models if their accuracy is higher than the accuracy property value of the currently registered model.

- E. Specify a tag named accuracy with the accuracy metric as a value when registering the model, and only register subsequent models if their accuracy is higher than the accuracy tag value of the currently registered model.

Correct Answer: *DE*
E: Using tags, you can track useful information such as the name and version of the machine learning library used to train the model. Note that tags must be alphanumeric.
Reference:
https://notebooks.azure.com/xavierheriat/projects/azureml-getting-started/html/how-to-use-azureml/deployment/register-model-create-image-deploy-service/ register-model-create-image-deploy-service.ipynb

Question #23*Topic 4*

You are a data scientist working for a hotel booking website company. You use the Azure Machine Learning service to train a model that identifies fraudulent transactions.
You must deploy the model as an Azure Machine Learning real-time web service using the Model.deploy method in the Azure Machine Learning SDK. The deployed web service must return real-time predictions of fraud based on transaction data input.
You need to create the script that is specified as the entry_script parameter for the InferenceConfig class used to deploy the model.
What should the entry script do?

- A. Register the model with appropriate tags and properties.
- B. Create a Conda environment for the web service compute and install the necessary Python packages.
- C. Load the model and use it to predict labels from input data.
- D. Start a node on the inference cluster where the web service is deployed.
- E. Specify the number of cores and the amount of memory required for the inference compute.

Correct Answer: *C*
The entry script receives data submitted to a deployed web service and passes it to the model. It then takes the response returned by the model and returns that to the client. The script is specific to your model. It must understand the data that the model expects and returns.
The two things you need to accomplish in your entry script are:
➮ Loading your model (using a function called init())
➮ Running your model on input data (using a function called run())
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-deploy-and-where

Question #24*Topic 4*

DRAG DROP -
You use Azure Machine Learning to deploy a model as a real-time web service.
You need to create an entry script for the service that ensures that the model is loaded when the service starts and is used to score new data as it is received.
Which functions should you include in the script? To answer, drag the appropriate functions to the correct actions. Each function 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.
Select and Place:

Answer Area

Functions	Action	Function
main()		
score()	Load the model when the service starts.	
run()	Use the model to score new data.	
init()		
predict()		

Correct Answer:

Answer Area

Functions	Action	Function
main()		
score()	Load the model when the service starts.	init()
run()	Use the model to score new data.	run()
init()		
predict()		

Box 1: init()
The entry script has only two required functions, init() and run(data). These functions are used to initialize the service at startup and run the model using request data passed in by a client. The rest of the script handles loading and running the model(s).
Box 2: run()
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-deploy-existing-model
Question #25*Topic 4*

You develop and train a machine learning model to predict fraudulent transactions for a hotel booking website.

Traffic to the site varies considerably. The site experiences heavy traffic on Monday and Friday and much lower traffic on other days. Holidays are also high web traffic days.
You need to deploy the model as an Azure Machine Learning real-time web service endpoint on compute that can dynamically scale up and down to support demand.
Which deployment compute option should you use?

- A. attached Azure Databricks cluster
- B. Azure Container Instance (ACI)
- C. Azure Kubernetes Service (AKS) inference cluster
- D. Azure Machine Learning Compute Instance
- E. attached virtual machine in a different region

Correct Answer: *C*
Azure Machine Learning compute cluster is a managed-compute infrastructure that allows you to easily create a single or multi-node compute. The compute is created within your workspace region as a resource that can be shared with other users in your workspace. The compute scales up automatically when a job is submitted, and can be put in an Azure Virtual Network.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-create-attach-compute-sdk

Question #26*Topic 4*

You use the designer to create a training pipeline for a classification model. The pipeline uses a dataset that includes the features and labels required for model training.
You create a real-time inference pipeline from the training pipeline. You observe that the schema for the generated web service input is based on the dataset and includes the label column that the model predicts. Client applications that use the service must not be required to submit this value.
You need to modify the inference pipeline to meet the requirement.
What should you do?

- A. Add a Select Columns in Dataset module to the inference pipeline after the dataset and use it to select all columns other than the label.
- B. Delete the dataset from the training pipeline and recreate the real-time inference pipeline.
- C. Delete the Web Service Input module from the inference pipeline.
- D. Replace the dataset in the inference pipeline with an Enter Data Manually module that includes data for the feature columns but not the label column.

Correct Answer: *A*
By default, the Web Service Input will expect the same data schema as the module output data which connects to the same downstream port as it. You can remove the target variable column in the inference pipeline using Select Columns in Dataset module. Make sure that the output of Select Columns in Dataset removing target variable column is connected to the same port as the output of the Web Service Intput module.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/tutorial-designer-automobile-price-deploy

Question #27*Topic 4*

You use the Azure Machine Learning designer to create and run a training pipeline. You then create a real-time inference pipeline.
You must deploy the real-time inference pipeline as a web service.
What must you do before you deploy the real-time inference pipeline?

- A. Run the real-time inference pipeline.
- B. Create a batch inference pipeline.
- C. Clone the training pipeline.
- D. Create an Azure Machine Learning compute cluster.

Correct Answer: *D*
You need to create an inferencing cluster.

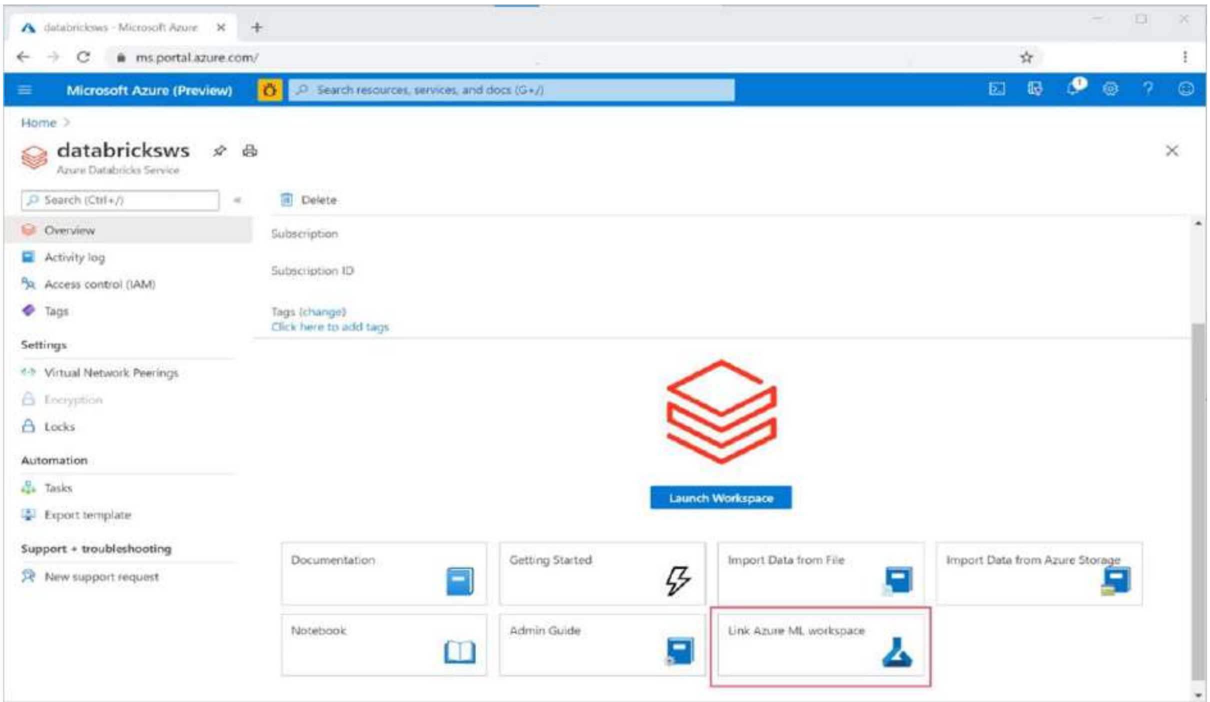
Deploy the real-time endpoint -
After your AKS service has finished provisioning, return to the real-time inferencing pipeline to complete deployment.
1. Select Deploy above the canvas.
2. Select Deploy new real-time endpoint.
3. Select the AKS cluster you created.
4. Select Deploy.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/tutorial-designer-automobile-price-deploy

Question #28*Topic 4*

You create an Azure Machine Learning workspace named ML-workspace. You also create an Azure Databricks workspace named DB-workspace. DB-workspace contains a cluster named DB-cluster.
You must use DB-cluster to run experiments from notebooks that you import into DB-workspace.
You need to use ML-workspace to track MLflow metrics and artifacts generated by experiments running on DB-cluster. The solution must minimize the need for custom code.
What should you do?

- A. From DB-cluster, configure the Advanced Logging option.
- B. From DB-workspace, configure the Link Azure ML workspace option.
- C. From ML-workspace, create an attached compute.
- D. From ML-workspace, create a compute cluster.

Correct Answer: *B*
Connect your Azure Databricks and Azure Machine Learning workspaces:
Linking your ADB workspace to your Azure Machine Learning workspace enables you to track your experiment data in the Azure Machine Learning workspace.
To link your ADB workspace to a new or existing Azure Machine Learning workspace
1. Sign in to Azure portal.
2. Navigate to your ADB workspace's Overview page.
3. Select the Link Azure Machine Learning workspace button on the bottom right.



Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-use-mlflow-azure-databricks
Question #29*Topic 4*

HOTSPOT -
You create an Azure Machine Learning workspace.
You need to detect data drift between a baseline dataset and a subsequent target dataset by using the DataDriftDetector class.
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.
Hot Area:

Answer Area

```
from azureml.core import Workspace, Dataset
from datetime import datetime

ws = Workspace.from_config()
dset = Dataset.get_by_name(ws, 'target')
baseline = target.time_before(datetime(2021, 2, 1))
features = ['windAngle', 'windSpeed', 'temperature', 'stationName']

monitor = DataDriftDetector.[ ▼ ] (ws, 'drift-monitor', baseline,
                            backfill
                            create_from_datasets
                            create_from_model
target, compute_target='cpu-cluster', frequency='Week', feature_list=None,
drift_threshold=.6, latency=24)

monitor = DataDriftDetector.get_by_name(ws, 'drift-monitor')
monitor = monitor.update(feature_list=features)
complete = monitor.[ ▼ ] (datetime(2021, 1, 1), datetime.today())
                  backfill
                  list
                  update
```

Correct Answer:

Answer Area

```
from azureml.core import Workspace, Dataset
from datetime import datetime

ws = Workspace.from_config()
dset = Dataset.get_by_name(ws, 'target')
baseline = target.time_before(datetime(2021, 2, 1))
features = ['windAngle', 'windSpeed', 'temperature', 'stationName']

monitor = DataDriftDetector. [create_from_datasets] (ws, 'drift-monitor', baseline,
target, compute_target='cpu-cluster', frequency='Week', feature_list=None,
drift_threshold=.6, latency=24)

monitor = DataDriftDetector.get_by_name(ws, 'drift-monitor')
monitor = monitor.update(feature_list=features)
complete = monitor. [backfill] (datetime(2021, 1, 1), datetime.today())
```

Box 1 options: backfill / create_from_datasets (selected) / create_from_model

Box 2 options: backfill (selected) / list / update

Box 1: create_from_datasets -
The create_from_datasets method creates a new DataDriftDetector object from a baseline tabular dataset and a target time series dataset.

Box 2: backfill -
The backfill method runs a backfill job over a given specified start and end date.
Syntax: backfill(start_date, end_date, compute_target=None, create_compute_target=False)
Incorrect Answers:
List and update do not have datetime parameters.
Reference:
https://docs.microsoft.com/en-us/python/api/azureml-datadrift/azureml.datadrift.datadriftdetector(class)
Question #30*Topic 4*

You are planning to register a trained model in an Azure Machine Learning workspace.
You must store additional metadata about the model in a key-value format. You must be able to add new metadata and modify or delete metadata after creation.
You need to register the model.
Which parameter should you use?

- A. description
- B. model_framework
- C. tags
- D. properties

Correct Answer: C
azureml.core.Model.properties:
Dictionary of key value properties for the Model. These properties cannot be changed after

registration, however new key value pairs can be added.
Reference:
https://docs.microsoft.com/en-us/python/api/azureml-core/azureml.core.model.model

Question #31*Topic 4*

You have a Python script that executes a pipeline. The script includes the following code: from azureml.core import Experiment pipeline_run = Experiment(ws, 'pipeline_test').submit(pipeline)
You want to test the pipeline before deploying the script.
You need to display the pipeline run details written to the STDOUT output when the pipeline completes.
Which code segment should you add to the test script?

- A. pipeline_run.get_metrics()
- B. pipeline_run.wait_for_completion(show_output=True)
- C. pipeline_param = PipelineParameter(name="stdout", default_value="console")
- D. pipeline_run.get_status()

Correct Answer: *B*
wait_for_completion: Wait for the completion of this run. Returns the status object after the wait.
Syntax: wait_for_completion(show_output=False, wait_post_processing=False, raise_on_error=True)

Parameter: show_output -
Indicates whether to show the run output on sys.stdout.

Question #32*Topic 4*

You train and register a machine learning model. You create a batch inference pipeline that uses the model to generate predictions from multiple data files.
You must publish the batch inference pipeline as a service that can be scheduled to run every night.
You need to select an appropriate compute target for the inference service.
Which compute target should you use?

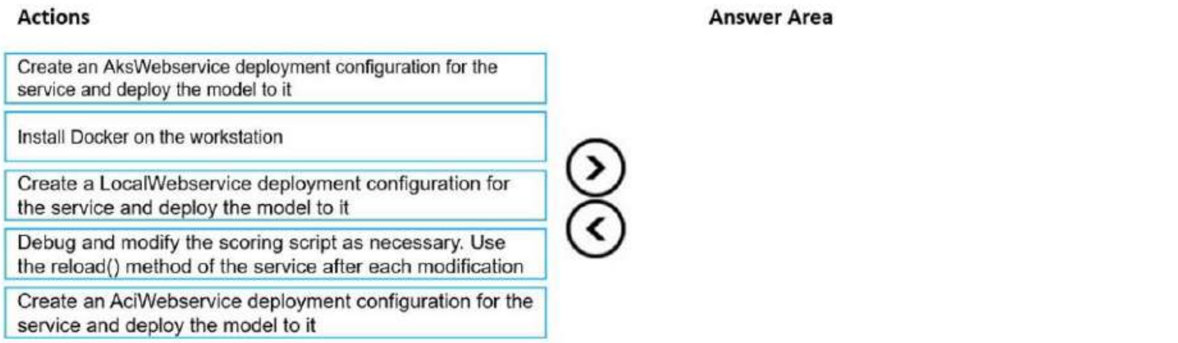
- A. Azure Machine Learning compute instance
- B. Azure Machine Learning compute cluster
- C. Azure Kubernetes Service (AKS)-based inference cluster
- D. Azure Container Instance (ACI) compute target

Correct Answer: *C*
Azure Machine Learning compute clusters is used for Batch inference. Run batch scoring on serverless compute. Supports normal and low-priority VMs. No support for real-time inference.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/concept-compute-target

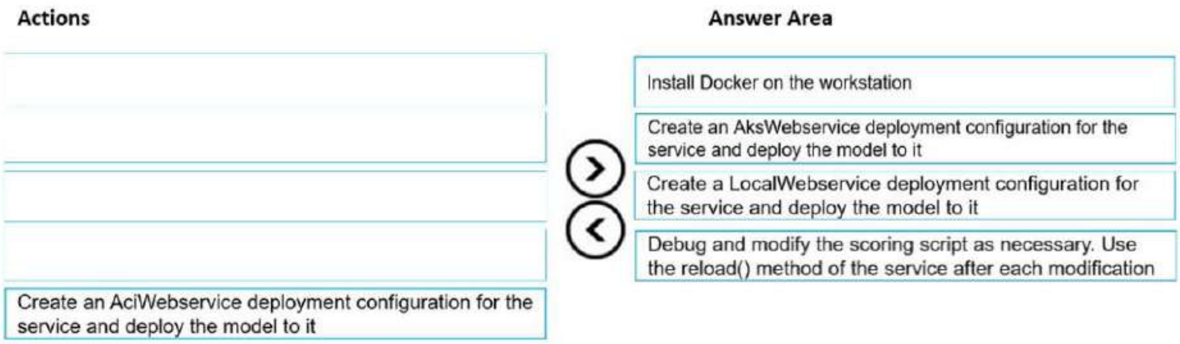
Question #33*Topic 4*

DRAG DROP -
You train and register a model by using the Azure Machine Learning SDK on a local workstation.

Python 3.6 and Visual Studio Code are installed on the workstation.
When you try to deploy the model into production as an Azure Kubernetes Service (AKS)-based web service, you experience an error in the scoring script that causes deployment to fail.
You need to debug the service on the local workstation before deploying the service to production.
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.
Select and Place:



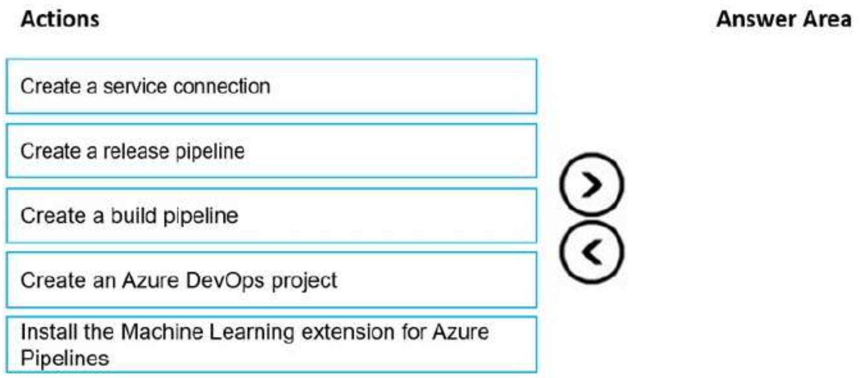
Correct Answer:



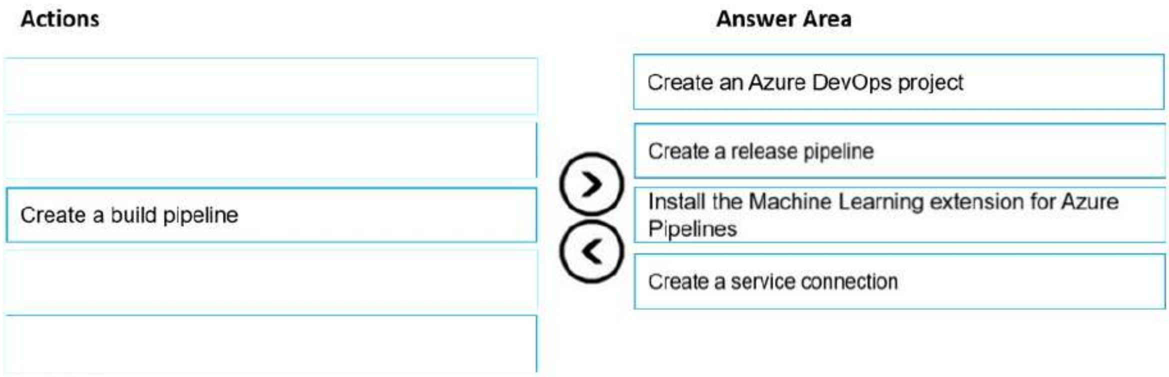
Step 1: Install Docker on the workstation
Prerequisites include having a working Docker installation on your local system.
Build or download the dockerfile to the compute node.
Step 2: Create an AksWebservice deployment configuration and deploy the model to it
To deploy a model to Azure Kubernetes Service, create a deployment configuration that describes the compute resources needed.
# If deploying to a cluster configured for dev/test, ensure that it was created with enough
# cores and memory to handle this deployment configuration. Note that memory is also used by
# things such as dependencies and AML components.
deployment_config = AksWebservice.deploy_configuration(cpu_cores = 1, memory_gb = 1) service = Model.deploy(ws, "myservice", [model], inference_config, deployment_config, aks_target)
service.wait_for_deployment(show_output = True) print(service.state) print(service.get_logs())
Step 3: Create a LocalWebservice deployment configuration for the service and deploy the model to it
To deploy locally, modify your code to use LocalWebservice.deploy_configuration() to create a deployment configuration. Then use Model.deploy() to deploy the service.
Step 4: Debug and modify the scoring script as necessary. Use the reload() method of the service after each modification.
During local testing, you may need to update the score.py file to add logging or attempt to resolve any problems that you've discovered. To reload changes to the score.py file, use reload(). For example, the following code reloads the script for the service, and then sends data to it.

Incorrect Answers:
✑ AciWebservice: The types of web services that can be deployed are LocalWebservice, which will deploy a model locally, and AciWebservice and
AksWebservice, which will deploy a model to Azure Container Instances (ACI) and Azure Kubernetes Service (AKS), respectively.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-deploy-azure-kubernetes-service
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-troubleshoot-deployment-local
Question #34*Topic 4*

DRAG DROP -
You create an Azure Machine Learning workspace and a new Azure DevOps organization. You register a model in the workspace and deploy the model to the target environment.
All new versions of the model registered in the workspace must automatically be deployed to the target environment.
You need to configure Azure Pipelines to deploy the model.
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.
Select and Place:

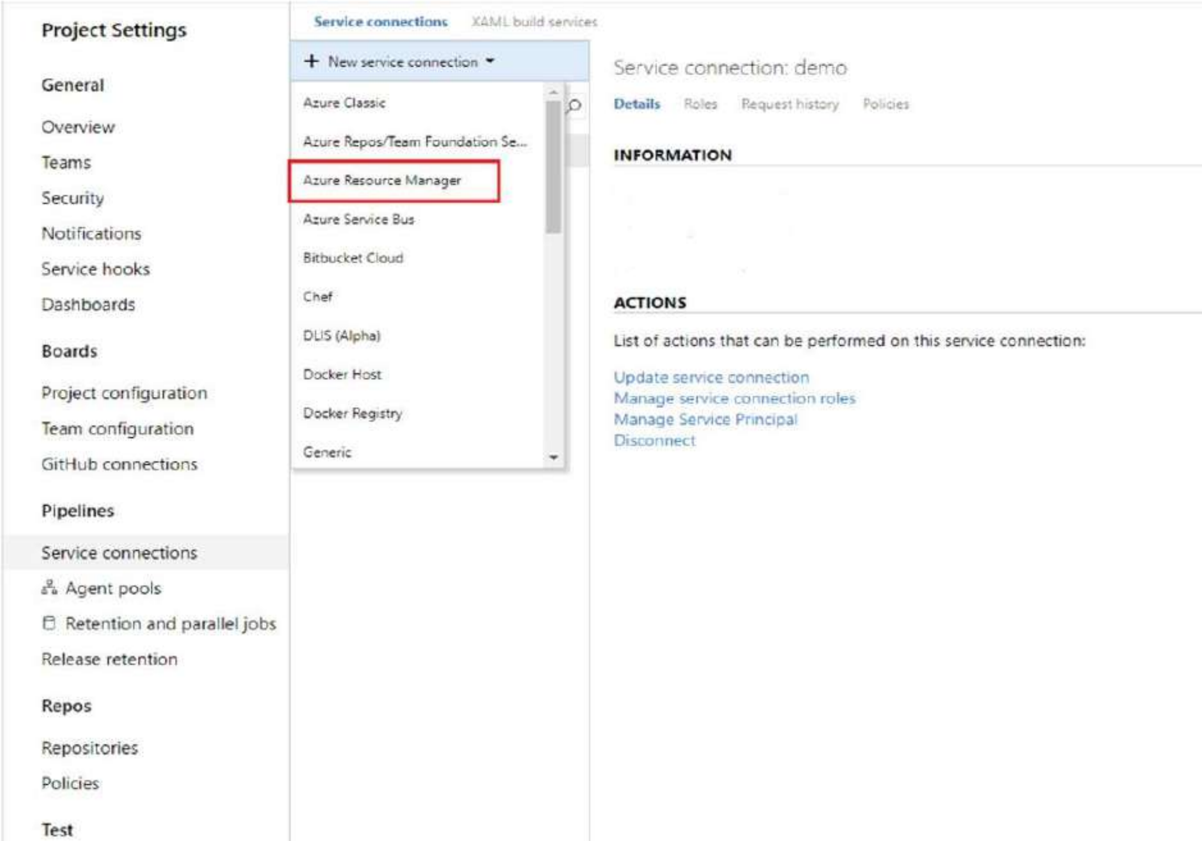


Correct Answer:

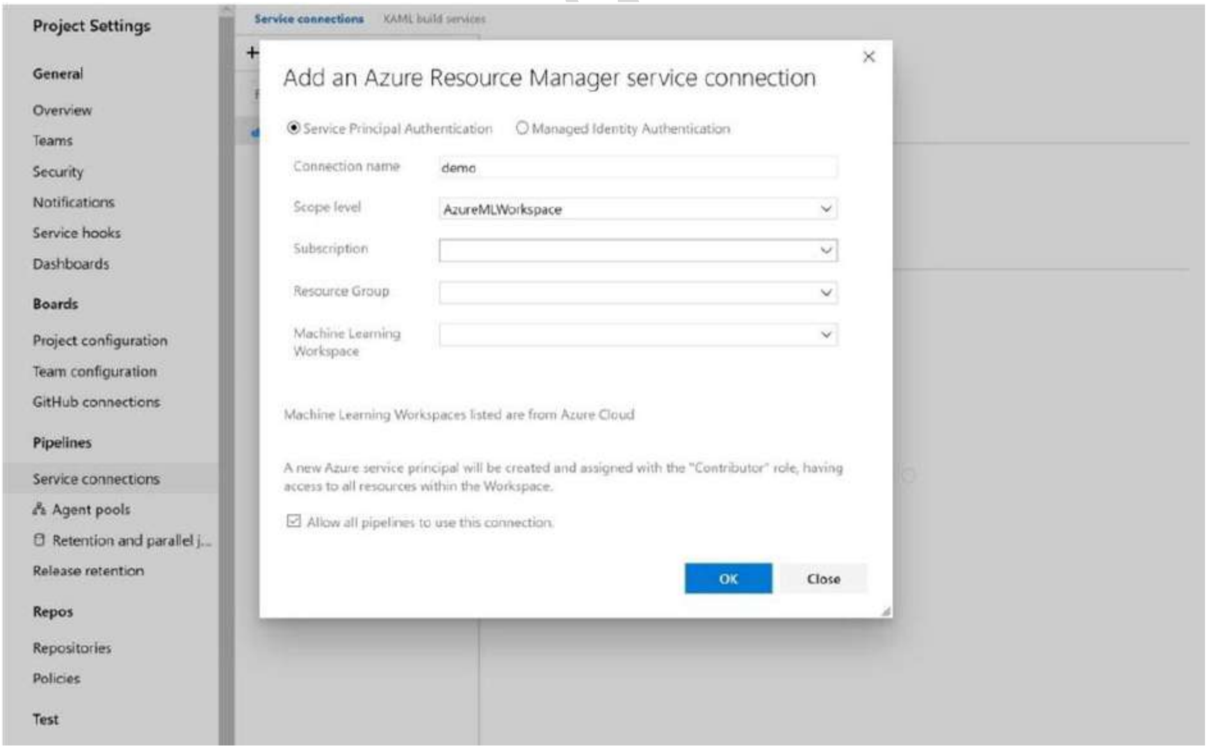


Step 1: Create an Azure DevOps project
Step 2: Create a release pipeline
1. Sign in to your Azure DevOps organization and navigate to your project.
2. Go to Pipelines, and then select New pipeline.
Step 3: Install the Machine Learning extension for Azure Pipelines
You must install and configure the Azure CLI and ML extension.

Step 4: Create a service connection
How to set up your service connection



Select AzureMLWorkspace for the scope level, then fill in the following subsequent parameters.



Note: How to enable model triggering in a release pipeline

➾ Go to your release pipeline and add a new artifact. Click on AzureML Model artifact then select the appropriate AzureML service connection and select from the available models in your workspace.

➩ Enable the deployment trigger on your model artifact as shown here. Every time a new version of that model is registered, a release pipeline will be triggered.
Reference:
https://marketplace.visualstudio.com/items?itemName=ms-air-aiagility.vss-services-azureml
https://docs.microsoft.com/en-us/azure/devops/pipelines/targets/azure-machine-learning
Question #35*Topic 4*

You use the Azure Machine Learning designer to create and run a training pipeline.
The pipeline must be run every night to inference predictions from a large volume of files. The folder where the files will be stored is defined as a dataset.
You need to publish the pipeline as a REST service that can be used for the nightly inferencing run.
What should you do?

- A. Create a batch inference pipeline
- B. Set the compute target for the pipeline to an inference cluster
- C. Create a real-time inference pipeline
- D. Clone the pipeline

Correct Answer: *A*
Azure Machine Learning Batch Inference targets large inference jobs that are not time-sensitive. Batch Inference provides cost-effective inference compute scaling, with unparalleled throughput for asynchronous applications. It is optimized for high-throughput, fire-and-forget inference over large collections of data.
You can submit a batch inference job by pipeline_run, or through REST calls with a published pipeline.
Reference:
https://github.com/Azure/MachineLearningNotebooks/blob/master/how-to-use-azureml/machine-learning-pipelines/parallel-run/README.md

Question #36*Topic 4*

HOTSPOT
-

You create an Azure Machine Learning model to include model files and a scoring script.

You must deploy the model. The deployment solution must meet the following requirements:

• Provide near real-time inferencing.
• Enable endpoint and deployment level cost estimates.
• Support logging to Azure Log Analytics.

You need to configure the deployment solution.

What should you configure? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Answer Area

Requirement | **Value**

Endpoint type

- Managed online
- Kubernetes online
- Batch

Deployment component

- Docker image
- Azure Container Instances (ACI)
- Azure Kubernetes Service (AKS) cluster

Correct Answer:

Answer Area

Requirement | **Value**

Endpoint type

- Managed online
- Kubernetes online
- Batch

Deployment component

- Docker image
- Azure Container Instances (ACI)
- Azure Kubernetes Service (AKS) cluster

Question #37*Topic 4*

You are developing a machine learning model.

You must inference the machine learning model for testing.

You need to use a minimal cost compute target.

Which two compute targets should you use? Each correct answer presents a complete solution.

NOTE: Each correct selection is worth one point.

- A. Azure Machine Learning Kubernetes
- B. Azure Databricks
- C. Remote VM
- D. Local web service
- E. Azure Container Instances

Correct Answer: *DE*

Question #38*Topic 4*

You train and publish a machine learning model.

You need to run a pipeline that retrains the model based on a trigger from an external system.

What should you configure?

- A. Azure Data Catalog
- B. Azure Batch
- C. Azure Logic App

Correct Answer: *C*

Question #39*Topic 4*

You create an Azure Machine Learning workspace.

You must configure an event handler to send an email notification when data drift is detected in the workspace datasets. You must minimize development efforts.

You need to configure an Azure service to send the notification.

Which Azure service should you use?

- A. Azure Logic Apps
- B. Azure Automation runbook
- C. Azure Function apps
- D. Azure DevOps pipeline

Correct Answer: *A*

Question #40*Topic 4*

HOTSPOT

-

You create an Azure Machine Learning dataset. You use the Azure Machine Learning designer to transform the dataset by using an Execute Python Script component and custom code.

You must upload the script and associated libraries as a script bundle.

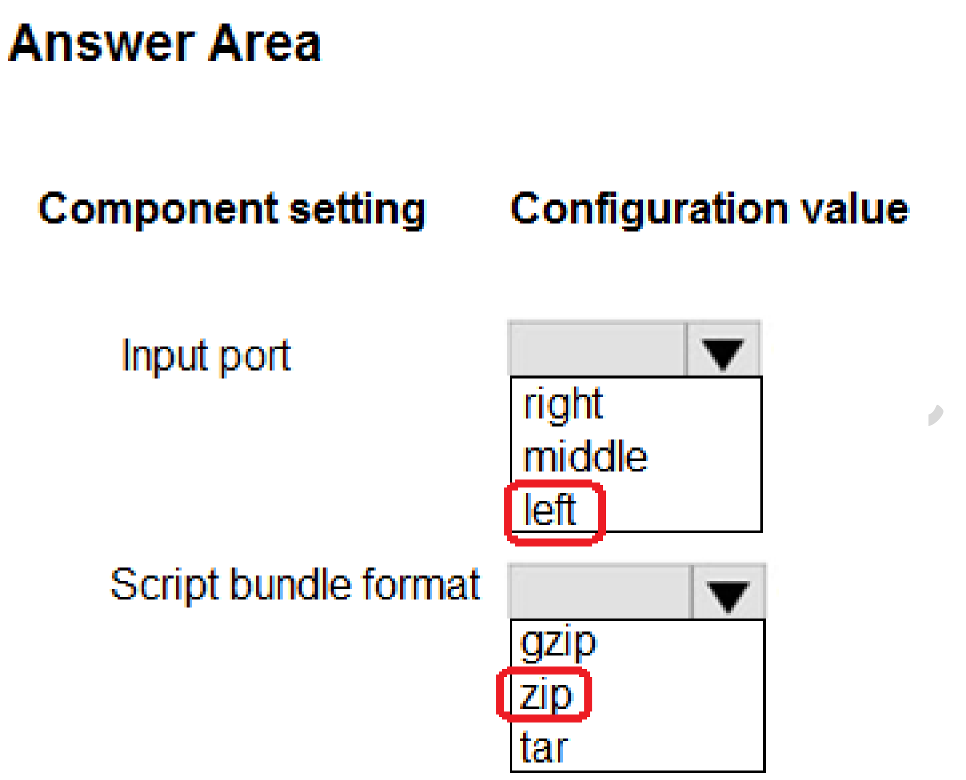
You need to configure the Execute Python Script component.

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

NOTE: Each correct selection is worth one point.

Answer Area

Component setting	Configuration value
Input port	right middle left
Script bundle format	gzip zip tar



Correct Answer:

Question #41*Topic 4*

HOTSPOT

-

You create a list of movie descriptions in text data format.

You must analyze the movie descriptions with automated machine learning.

You need to use the Azure Machine Learning for Python SDK v1 to configure a job with the specific natural language processing (NLP) task function for AutoML jobs.

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

NOTE: Each correct selection is worth one point.

Answer Area

Requirement	Function
Classify a movie description as either comedy or romantic.	text_classification() text_classification_multilable() text_ner()
Classify a movie description as either comedy, romantic, or both comedy and romantic.	text_classification() text_classification_multilable() text_ner()
Extract locations such as London or Paris from a movie description.	text_classification() text_classification_multilable() text_ner()

Correct Answer:

Answer Area

Requirement	Function
Classify a movie description as either comedy or romantic.	text_classification() text_classification_multilable() text_ner()
Classify a movie description as either comedy, romantic, or both comedy and romantic.	text_classification() text_classification_multilable() text_ner()
Extract locations such as London or Paris from a movie description.	text_classification() text_classification_multilable() text_ner()

Question #42*Topic 4*

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 create an Azure Machine Learning pipeline named pipeline1 with two steps that contain Python scripts. Data processed by the first step is passed to the second step.

You must update the content of the downstream data source of pipeline1 and run the pipeline again.

You need to ensure the new run of pipeline1 fully processes the updated content.

Solution: Set the allow_reuse parameter of the PythonScriptStep object of both steps to False.

Does the solution meet the goal?

- A. Yes
- B. No

Correct Answer: *A*

Question #43*Topic 4*

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 create an Azure Machine Learning pipeline named pipeline1 with two steps that contain Python scripts. Data processed by the first step is passed to the second step.

You must update the content of the downstream data source of pipeline1 and run the pipeline again.

You need to ensure the new run of pipeline1 fully processes the updated content.

Solution: Set the regenerate_outputs parameter of the pipeline1 experiment's run submit method to True.

Does the solution meet the goal?

- A. Yes
- B. No

Correct Answer: *A*

Question #44*Topic 4*

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 create an Azure Machine Learning pipeline named pipeline1 with two steps that contain Python scripts. Data processed by the first step is passed to the second step.

You must update the content of the downstream data source of pipeline1 and run the pipeline again.

You need to ensure the new run of pipeline1 fully processes the updated content.

Solution: Change the value of the compute_target parameter of the PythonScriptStep object in the two steps.

Does the solution meet the goal?

- A. Yes
- B. No

Correct Answer: *B*

Question #45*Topic 4*

HOTSPOT

-

You are authoring a pipeline by using the Azure Machine Learning SDK for Python. You implement code to import all relevant classes, configure the workspace, and define all pipeline steps.

You need to initiate pipeline execution.

How should you complete the code? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Correct Answer:

Question #46*Topic 4*

DRAG DROP

-

You have an Azure Machine Learning workspace that contains a training cluster and an inference cluster.

You plan to create a classification model by using the Azure Machine Learning designer.

You need to ensure that client applications can submit data as HTTP requests and receive predictions as responses.

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 area

Deploy a service to the compute cluster.

Create a pipeline that trains a classification model and run the pipeline on the compute cluster.

Create a real-time inference pipeline and run the pipeline on the compute cluster.

Correct Answer:

Question #47*Topic 4*

You create an MLflow model.

You must deploy the model to Azure Machine Learning for batch inference.

You need to create the batch deployment.

Which two components should you use? Each correct answer presents a complete solution.

NOTE: Each correct selection is worth one point.

- A. Environment
- B. Model files
- C. Online endpoint
- D. Kubernetes online endpoint
- E. Compute target

Correct Answer: *BE*

Question #48*Topic 4*

You create an Azure Machine Learning workspace. The workspace contains a dataset named sample_dataset, a compute instance, and a compute cluster.

You must create a two-stage pipeline that will prepare data in the dataset and then train and register a model based on the prepared data.

The first stage of the pipeline contains the following code:

```
from azureml.data import OutputFileDatasetConfig
from azureml.pipeline.steps import PythonScriptStep

sample_dataset = ws.datasets.get("sample_dataset")
stage1_data = OutputFileDatasetConfig("stage1_data")
stage1_step = PythonScriptStep(name = "stage1",
                        source_directory = 'source_data_container',
                        script_name = "stage1_script.py",
                        arguments = ['--input-data', sample_dataset.as_named_input('raw_data'),
                             '--prepped data', stage1_data]
                        compute_target = compute_cluster,
                        runconfig = pipeline_run_config,
                        allow_reuse = True)
```

You need to identify the location containing the output of the first stage of the script that you can use as input for the second stage.

Which storage location should you use?

- A. workspaceblobstore datastore
- B. workspacefilestore datastore
- C. compute instance
- D. compute_cluster

Correct Answer: *A*

Question #49*Topic 4*

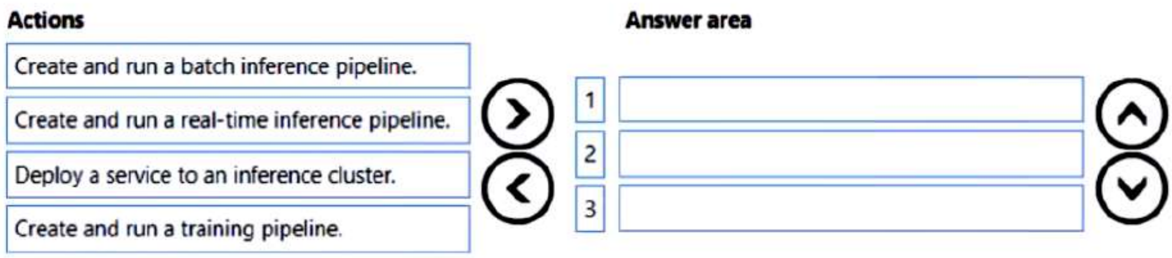
DRAG DROP

-

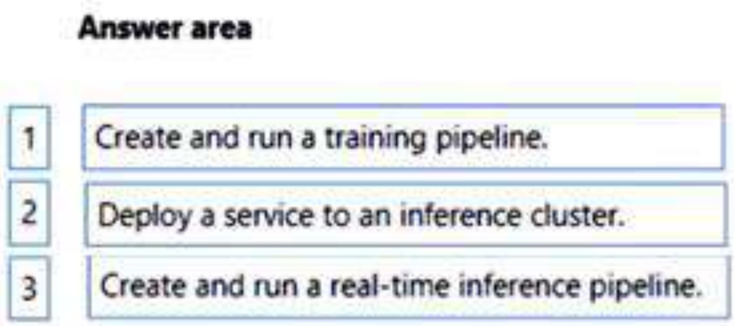
You are developing a machine learning solution by using the Azure Machine Learning designer.

You need to create a web service that applications can use to submit data feature values and retrieve a predicted label.

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.



Correct Answer:



Question #50*Topic 4*

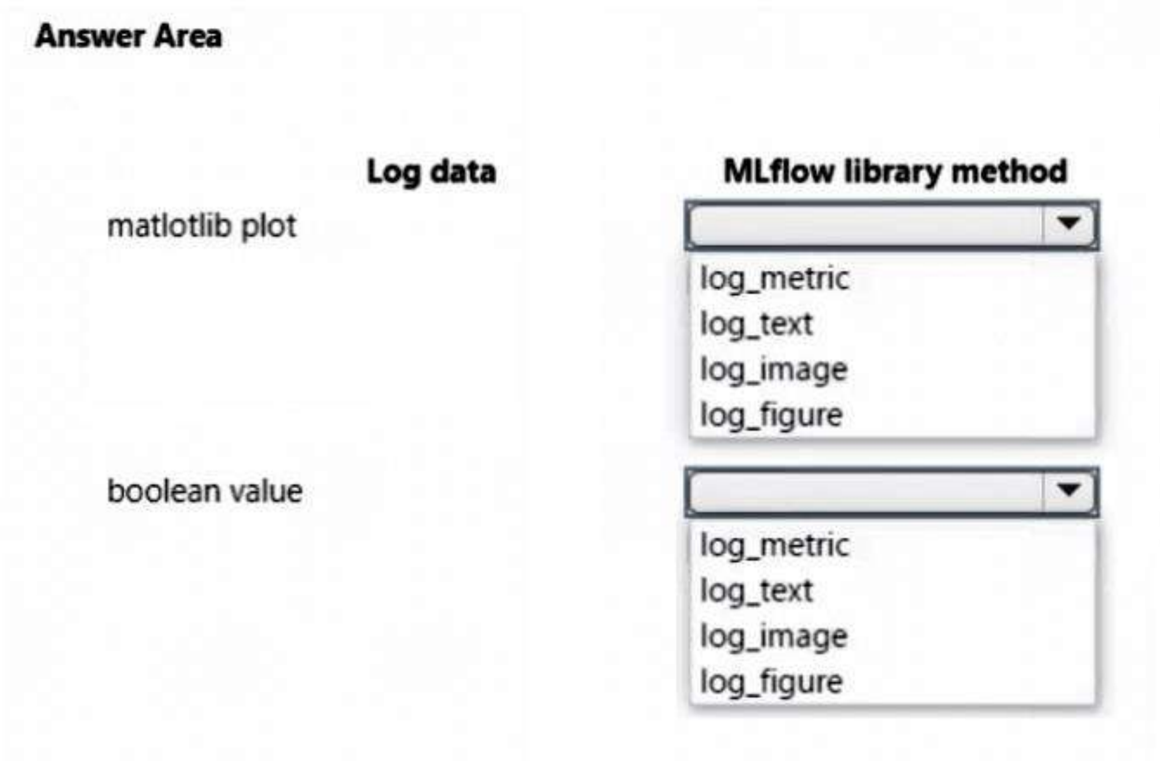
HOTSPOT
-

You create an Azure Machine Learning workspace and install the MLflow library.

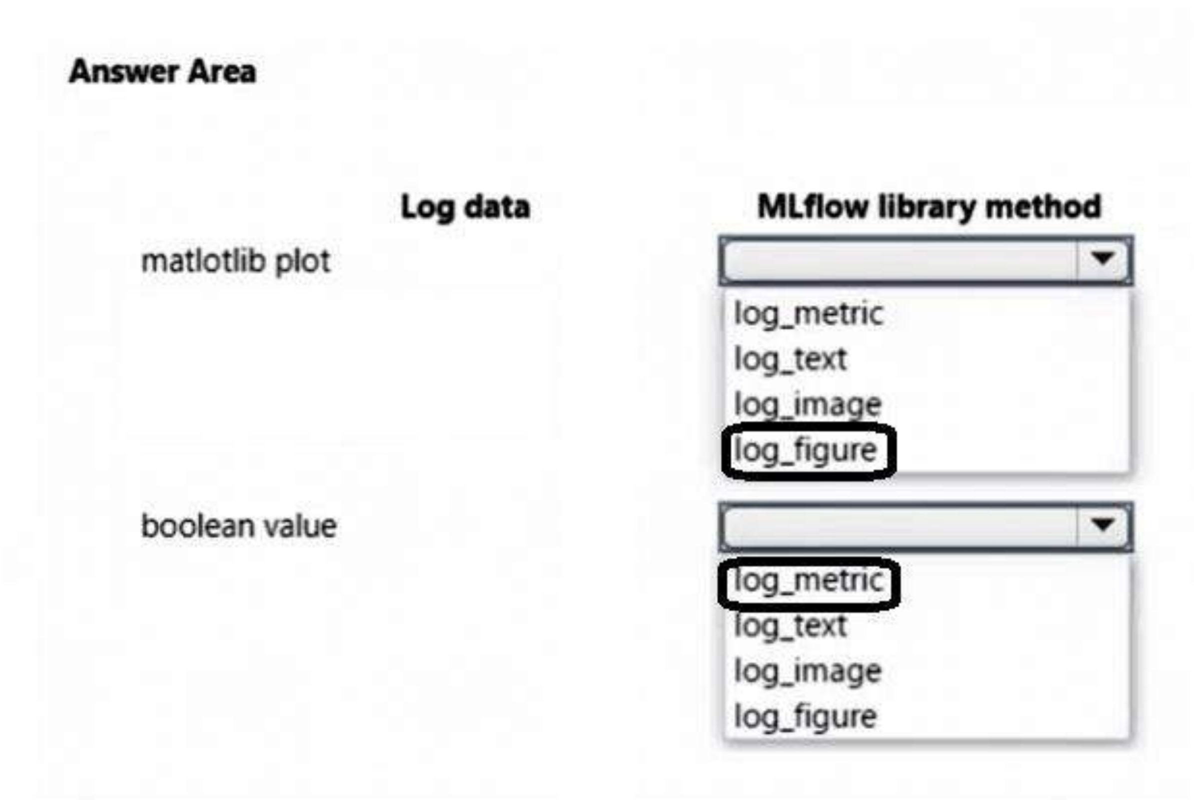
You need to log different types of data by using the MLflow library.

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

NOTE: Each correct selection is worth one point.



Correct Answer:



Question #51*Topic 4*

DRAG DROP

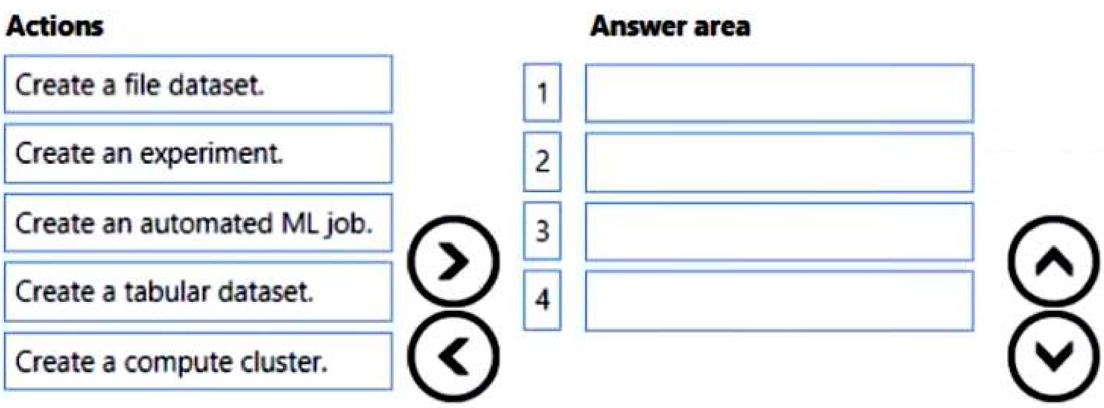
-

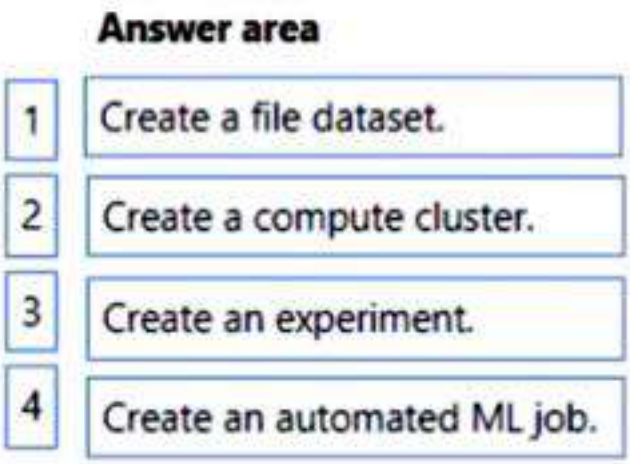
You create an Azure Machine Learning workspace. You are training a classification model with no-code AutoML in Azure Machine Learning studio.

The model must predict if a client of a financial institution will subscribe to a fixed-term deposit. You must preview the data profile in Azure Machine Learning studio once the dataset is created.

You need to train the model.

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.





Correct Answer:

Question #52*Topic 4*

You create an Azure Machine Learning workspace. You use Azure Machine Learning designer to create a pipeline within the workspace.

You need to submit a pipeline run from the designer.

What should you do first?

- A. Create an experiment.
- B. Create an attached compute resource.
- C. Create a compute cluster.
- D. Select a model.

Correct Answer: *C*

Question #53*Topic 4*

HOTSPOT

-

You create an Azure Machine learning workspace. The workspace contains a folder named src. The folder contains a Python script named script1.py.

You use the Azure Machine Learning Python SDK v2 to create a control script. You must use the control script to run script1.py as part of a training job.

You need to complete the section of script that defines the job parameters.

How should you complete the script? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Answer Area

```
ws = Workspace.from_config()
ml_client = MLClient(
    ws.subscription_id,
    ws.resource_group,
    ws.name)
job = [ ▼ ](
      command
      mlclient
      ws
[ ▼ ]="./src",
 code
 inputs
 path
    command="python script1.py",
    environment="AzureML-sklearn-0.24-ubuntu18.04-py37-cpu@latest",
    compute="cpu-cluster",
    display_name="hello-world-example",
)
```

Correct Answer:

Answer Area

```
ws = Workspace.from_config()
ml_client = MLClient(
    ws.subscription_id,
    ws.resource_group,
    ws.name)
job = [command | mlclient | ws] (
[code | inputs | path] ="./src",
    command="python script1.py",
    environment="AzureML-sklearn-0.24-ubuntu18.04-py37-cpu@latest",
    compute="cpu-cluster",
    display_name="hello-world-example",
)
```

Question #54*Topic 4*

HOTSPOT

-

You manage an Azure Machine Learning workspace. You submit a training job with the Azure Machine Learning Python SDK v2. You must use MLflow to log metrics, model parameters, and model artifacts automatically when training a model.

You start by writing the following code segment:

import mlflow
mlflow.autolog(log_models=False, exclusive=True)

For each of the following statements, select Yes if the statement is true. Otherwise, select No.

Answer Area

Statements	Yes	No
The code enables logging of autologged content to a user-created fluent run.	○	○
Trained models are logged as MLflow model artifacts.	○	○
All metrics and parameters are logged during training.	○	○

Correct Answer:

Answer Area

Statements	Yes	No
The code enables logging of autologged content to a user-created fluent run.		
Trained models are logged as MLflow model artifacts.		
All metrics and parameters are logged during training.		

Question #55*Topic 4*

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 train and register an Azure Machine Learning model.

You plan to deploy the model to an online endpoint.

You need to ensure that applications will be able to use the authentication method with a non-expiring artifact to access the model.

Solution: Create a managed online endpoint and set the value of its auth_mode parameter to aml_token. Deploy the model to the online endpoint.

Does the solution meet the goal?

- A. Yes
- B. No

Correct Answer: *B*

Question #56*Topic 4*

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 train and register an Azure Machine Learning model.

You plan to deploy the model to an online endpoint.

You need to ensure that applications will be able to use the authentication method with a non-expiring artifact to access the model.

Solution: Create a managed online endpoint and set the value of its auth_mode parameter to key. Deploy the model to the online endpoint.

Does the solution meet the goal?

- A. Yes
- B. No

Correct Answer: *A*

Question #57*Topic 4*

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 train and register an Azure Machine Learning model.

You plan to deploy the model to an online endpoint.

You need to ensure that applications will be able to use the authentication method with a non-expiring artifact to access the model.

Solution: Create a managed online endpoint with the default authentication settings. Deploy the model to the online endpoint.

Does the solution meet the goal?

- A. Yes
- B. No

Correct Answer: *B*

Question #58*Topic 4*

You use the Azure Machine Learning SDK v2 for Python and notebooks to train a model. You use Python code to create a compute target, an environment, and a training script.

You need to prepare information to submit a training job.

Which class should you use?

- A. MLClient
- B. BuildContext
- C. EndpointConnection
- D. command

Correct Answer: *D*

Question #59*Topic 4*

You manage an Azure Machine Learning workspace.

You build a custom model you must log with MLflow. The custom model includes the following:

• The model is not natively supported by MLflow.
• The model cannot be serialized in Pickle format.
• The model source code is complex.
• The Python library for the model must be packaged with the model.

You need to create a custom model flavor to enable logging with MLflow.

What should you use?

- A. model loader
- B. artifacts
- C. model wrapper
- D. custom signatures

Correct Answer: *A*

Question #60*Topic 4*

You create an Azure Machine Learning workspace.

You must use the Python SDK v2 to implement an experiment from a Jupyter notebook in the workspace. The experiment must log string metrics.

You need to implement the method to log the string metrics.

Which method should you use?

- A. mlflow.log_artifact()
- B. mlflow.log.dict()
- C. mlflow.log_metric()
- D. mlflow.log_text()

Correct Answer: *D*

Question #61*Topic 4*

You manage an Azure Machine Learning workspace. You develop a machine learning model.

You must deploy the model to use a low-priority VM with a pricing discount.

You need to deploy the model.

Which compute target should you use?

- A. Azure Kubernetes Service (AKS)
- B. Azure Machine Learning compute clusters
- C. Azure Container Instances (ACI)
- D. Local deployment

Correct Answer: *B*

Question #62*Topic 4*

HOTSPOT

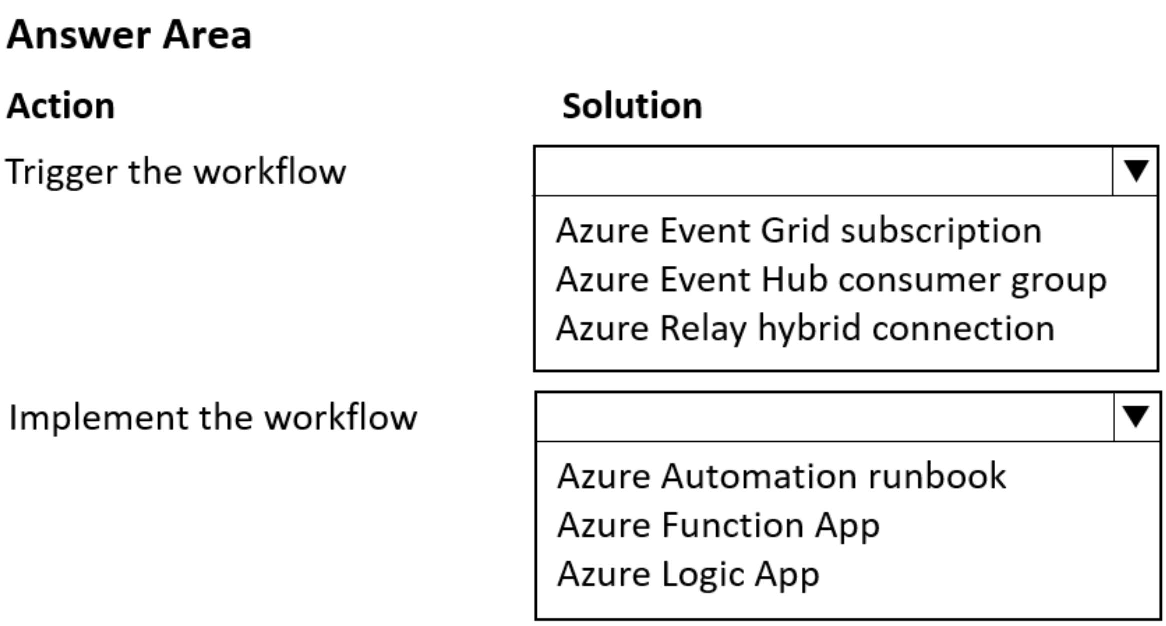
-

You use an Azure Machine Learning workspace, Azure Data Factory pipeline, and a dataset monitor that runs on a schedule to detect data drift.

You need to implement an automated workflow to trigger when the dataset monitor detects data drift and launch the Azure Data Factory pipeline to update the dataset. The solution must minimize the effort to configure the workflow.

How should you configure the workflow? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.



Correct Answer:

Answer Area

Action	Solution
Trigger the workflow	Azure Event Grid subscription Azure Event Hub consumer group Azure Relay hybrid connection
Implement the workflow	Azure Automation runbook Azure Function App Azure Logic App

Question #63*Topic 4*

DRAG DROP

-

You train and register a model by using the Azure Machine Learning Python SDK v2 on a local workstation. Python 3.7 and Visual Studio Code are installed on the workstation.

When you try to deploy the model into production to a Kubernetes online endpoint, you experience an error in the scoring script that causes deployment to fail.

You need to debug the service on the local workstation before deploying the service to production.

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.

Actions

- Install Docker on the workstation.
- Run the from.config method of an MLClient class instance with the path parameter set to local.
- Debug and modify the scoring script as necessary.
- Run the begin_create_or_update method of an MLClient class instance with the local parameter set to true.
- Upgrade the Python version on the workstation.

Answer Area

Answer Area

Install Docker on the workstation.

Run the begin_create_or_update method of an MLClient class instance with the local parameter set to true.

Debug and modify the scoring script as necessary.

Correct Answer:

Question #64*Topic 4*

DRAG DROP

-

You create a multi-class image classification deep learning model.

The model must be retrained monthly with the new image data fetched from a public web portal. You create an Azure Machine Learning pipeline to fetch new data, standardize the size of images, and retrain the model.

You need to use the Azure Machine Learning Python SDK v2 to configure the schedule for the pipeline. The schedule should be defined by using the frequency and interval properties, with frequency set to “month" and interval set to "1".

Which three classes should you instantiate in sequence? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.

Actions

- JobSchedule
- CronTrigger
- AutoPauseSettings
- PipelineJob
- RecurrenceTrigger

Answer area

Answer area

PipelineJob

RecurrenceTrigger

JobSchedule

Correct Answer:

Question #65*Topic 4*

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 train and register an Azure Machine Learning model.

You plan to deploy the model to an online endpoint.

You need to ensure that applications will be able to use the authentication method with a non-expiring artifact to access the model.

Solution: Create a Kubernetes online endpoint and set the value of its auth_mode parameter to aml_token. Deploy the model to the online endpoint.

Does the solution meet the goal?

- A. Yes
- B. No

Correct Answer: *B*

Question #66*Topic 4*

You manage an Azure Machine Learning workspace.

You must provide explanations for the behavior of the models with feature importance measures.

You need to configure a Responsible AI dashboard in Azure Machine Learning.

Which dashboard component should you configure?

- A. Counterfactual what-if
- B. Casual inference
- C. Fairness assessment
- D. Interpretability

Correct Answer: *D*

Question #67*Topic 4*

You create an Azure Machine Learning workspace.

You must use the Python SDK v2 to implement an experiment from a Jupyter notebook in the workspace. The experiment must log a list of numeral metrics.

You need to implement a method to log a list of numeral metrics.

Which method should you use?

- A. mlflow.log_metric()
- B. mlflow.log.batch()
- C. mlflow.log_image()
- D. mlflow.log_artifact()

Correct Answer: *A*

Question #68*Topic 4*

HOTSPOT
-

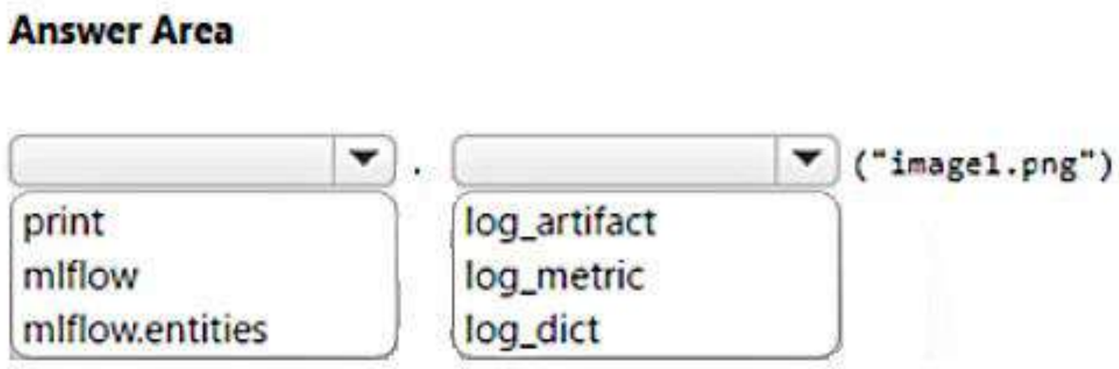
You create an Azure Machine Learning workspace.

You plan to write an Azure Machine Learning SDK for Python v2 script that logs an image for an experiment. The logged image must be available from the images tab in Azure Machine Learning Studio.

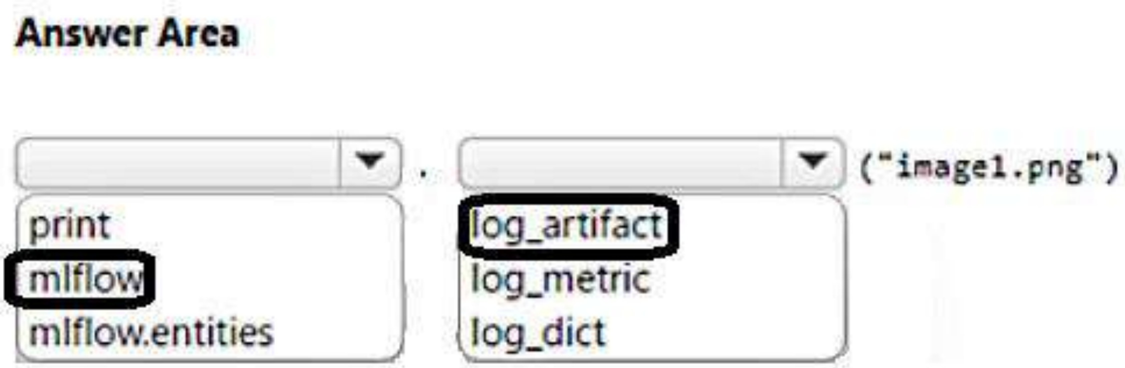
You need to complete the script.

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

NOTE: Each correct selection is worth one point.



Correct Answer:



Question #69*Topic 4*

You run Azure Machine Learning training experiments. The training scripts directory contains 100 files that includes a file named .amlignore. The directory also contains subdirectories named ./outputs and ./logs.

There are 20 files in the training scripts directory that must be excluded from the snapshot to the compute targets. You create a file named .gitignore in the root of the directory. You add the names of the 20 files to the .gitignore file. These 20 files continue to be copied to the compute targets.

You need to exclude the 20 files.

What should you do?

- A. Copy the contents of the file named .gitignore to the file named .amlignore.
- B. Move the file named .gitignore to the ./outputs directory.
- C. Move the file named .gitignore to the ./logs directory.
- D. Add the contents of the file named .amlignore to the file named .gitignore.

Correct Answer: *A*

Question #70*Topic 4*

You build a data pipeline in an Azure Machine Learning workspace by using the Azure Machine Learning SDK for Python.

You need to run a Python script as a pipeline step.

Which two classes could you use? Each correct answer presents a complete solution.

NOTE: Each correct selection is worth one point.

- A. PythonScriptStep
- B. AutoMLStep
- C. CommandStep
- D. StepRun

Correct Answer: *CD*

Question #71*Topic 4*
You have an Azure Machine Learning workspace.

You plan to run a job to train a model as an MLflow model output.

You need to specify the output mode of the MLflow model.

Which three modes can you specify? Each correct answer presents a complete solution.

NOTE: Each correct selection is worth one point.

- A. rw_mount
- B. ro_mount
- C. upload
- D. download
- E. direct

Correct Answer: *ACE*

Question #72*Topic 4*

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 have an Azure Machine Learning workspace. You connect to a terminal session from the Notebooks page in Azure Machine Learning studio.

You plan to add a new Jupyter kernel that will be accessible from the same terminal session.

You need to perform the task that must be completed before you can add the new kernel.

Solution: Create an environment.

Does the solution meet the goal?

- A. Yes
- B. No

Correct Answer: *A*

Question #73*Topic 4*

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 have an Azure Machine Learning workspace. You connect to a terminal session from the Notebooks page in Azure Machine Learning studio.

You plan to add a new Jupyter kernel that will be accessible from the same terminal session.

You need to perform the task that must be completed before you can add the new kernel.

Solution: Delete the Python 3.6 – AzureML kernel.

Does the solution meet the goal?

- A. Yes
- B. No

Correct Answer: *B*

Question #74*Topic 4*

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 have an Azure Machine Learning workspace. You connect to a terminal session from the Notebooks page in Azure Machine Learning studio.

You plan to add a new Jupyter kernel that will be accessible from the same terminal session.

You need to perform the task that must be completed before you can add the new kernel.

Solution: Delete the Python 3.8 – AzureML kernel.

Does the solution meet the goal?

- A. Yes
- B. No

Correct Answer: *B*

Question #75*Topic 4*

You are a data scientist working for a hotel booking website company. You use the Azure Machine Learning service to train a model that identifies fraudulent transactions.

You must deploy the model as an Azure Machine Learning online endpoint by using the Azure Machine Learning Python SDK v2. The deployed model must return real-time predictions of fraud based on transaction data input.

You need to create the script that is specified as the scoring_script parameter for the CodeConfiguration class used to deploy the model.

What should the entry script do?

- A. Register the model with appropriate tags and properties.
- B. Create a Conda environment for the online endpoint compute and install the necessary Python packages.
- C. Load the model and use it to predict labels from input data.
- D. Start a node on the inference cluster where the model is deployed.
- E. Specify the number of cores and the amount of memory required for the online endpoint compute.

Correct Answer: *C*

Question #76*Topic 4*

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 have an Azure Machine Learning workspace. You connect to a terminal session from the Notebooks page in Azure Machine Learning studio.

You plan to add a new Jupyter kernel that will be accessible from the same terminal session.

You need to perform the task that must be completed before you can add the new kernel.

Solution: Create a compute instance.

Does the solution meet the goal?

- A. Yes
- B. No

Correct Answer: *B*

Question #: 77
Topic #: 4

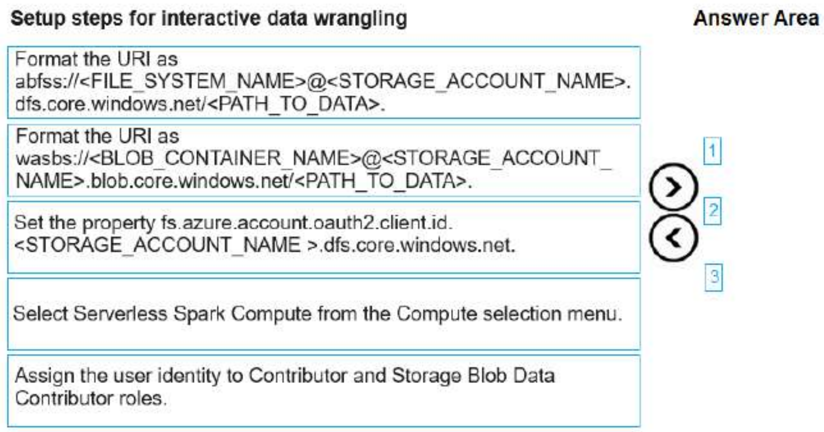
DRAG DROP
-

You manage an Azure Machine Learning workspace.

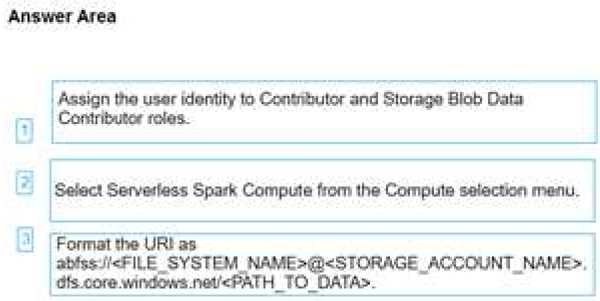
You plan to import and wrangle data stored in Azure Data Lake Storage Gen2 with Apache Spark.

You need to start interactive data wrangling with the user identity passthrough.

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.



Correct Answer:



Question #: 78
Topic #: 4

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 have an Azure Machine Learning workspace named Workspace1. Workspace1 has a registered MLflow model named model1 with PyFunc flavor.

You plan to deploy model1 to an online endpoint named endpoint1 without egress connectivity by using Azure Machine Learning Python SDK v2.

You have the following code:

```
blue_deployment = ManagedOnlineDeployment(
    name="blue",
    endpoint_name=endpoint1,
    model=model1,
    instance_type="Standard_F4s_v2",
    instance_count=1
)
```

You need to add a parameter to the ManagedOnlineDeployment object to ensure the model deploys successfully.

Solution: Add the scoring_script parameter.

Does the solution meet the goal?

- A. Yes
- B. No

Correct Answer: *A*

Question #: 79
Topic #: 4

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 have an Azure Machine Learning workspace named Workspace1. Workspace1 has a registered MLflow model named model1 with PyFunc flavor.

You plan to deploy model1 to an online endpoint named endpoint1 without egress connectivity by using Azure Machine Learning Python SDK v2.

You have the following code:

```
blue_deployment = ManagedOnlineDeployment(
    name="blue",
    endpoint_name=endpoint1,
    model=model1,
    instance_type="Standard_F4s_v2",
    instance_count=1
)
```

You need to add a parameter to the ManagedOnlineDeployment object to ensure the model deploys successfully.

Solution: Add the environment parameter.

Does the solution meet the goal?

- A. Yes
- B. No

Correct Answer: *B*

Question #: 80
Topic #: 4

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 have an Azure Machine Learning workspace named Workspace1. Workspace1 has a registered MLflow model named model1 with PyFunc flavor.

You plan to deploy model1 to an online endpoint named endpoint1 without egress connectivity by using Azure Machine Learning Python SDK v2.

You have the following code:

```
blue_deployment = ManagedOnlineDeployment(
    name="blue",
    endpoint_name=endpoint1,
    model=model1,
    instance_type="Standard_F4s_v2",
    instance_count=1
)
```

You need to add a parameter to the ManagedOnlineDeployment object to ensure the model deploys successfully.

Solution: Add the with_package parameter.

Does the solution meet the goal?

- A. Yes
- B. No

Correct Answer: *A*

Question #: 81
Topic #: 4

You manage an Azure Machine Learning workspace.

An MLflow model is already registered. You plan to customize how the deployment does inference.

You need to deploy the MLflow model to a batch endpoint for batch inferencing.

What should you create first?

- A. scoring script
- B. deployment
- C. environment
- D. deployment definition

Correct Answer: *D*

Question #: 82
Topic #: 4

You manage an Azure Machine Learning workspace that includes a batch endpoint.

You plan to deploy a model to the batch endpoint.

You need to configure compute for the deployment.

Which compute should you use?

- A. Remote VM
- B. AmlCompute instance
- C. Azure Batch
- D. Kubernetes cluster

Correct Answer: *B*

Question #: 83
Topic #: 4

You manage an Azure Machine Learning workspace.

You plan to train a natural language processing (NLP) text classification model in multiple languages by using Azure Machine Learning Python SDK v2.

You need to configure the language of the text classification job by using automated machine learning.

Which method of the TextClassificationJob class should you use?

- A. set_data
- B. set_featurization
- C. set_sweep
- D. set_training_parameters

Correct Answer: *B*

Question #: 84
Topic #: 4

DRAG DROP
-

You set up a machine learning workflow as an automated process. You have an Owner role in an Azure subscription that contains the Azure Machine Learning workspace.

You must set up an authentication method that allows an automated process to authenticate to the workspace without requiring user interaction.

You need to set up the authentication for the Azure Machine Learning workspace.

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.

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

Authentication steps

- Specify the tenant by explicitly instantiating InteractiveLoginAuthentication with Tenant ID as an argument.
- Create a service principal in Microsoft Entra ID by registering a new application.
- Add a client secret, write a description for your key, and select duration.
- Select Access Control (IAM) and grant the Contributor role for the service principal.
- Enable system-assigned managed identity to authenticate to the workspace.

Answer Area

Correct Answer:

Answer Area

- Enable system-assigned managed identity to authenticate to the workspace.
- Select Access Control (IAM) and grant the Contributor role for the service principal.
- Add a client secret, write a description for your key, and select duration.

Topic 5 - Question Set 5

Question #1*Topic 5*

You are a data scientist working for a bank and have used Azure ML to train and register a machine learning model that predicts whether a customer is likely to repay a loan.
You want to understand how your model is making selections and must be sure that the model does not violate government regulations such as denying loans based on where an applicant lives.
You need to determine the extent to which each feature in the customer data is influencing predictions.
What should you do?

- A. Enable data drift monitoring for the model and its training dataset.
- B. Score the model against some test data with known label values and use the results to calculate a confusion matrix.
- C. Use the Hyperdrive library to test the model with multiple hyperparameter values.

- D. Use the interpretability package to generate an explainer for the model.
- E. Add tags to the model registration indicating the names of the features in the training dataset.

Correct Answer: *D*
When you compute model explanations and visualize them, you're not limited to an existing model explanation for an automated ML model. You can also get an explanation for your model with different test data. The steps in this section show you how to compute and visualize engineered feature importance based on your test data.
Incorrect Answers:
A: In the context of machine learning, data drift is the change in model input data that leads to model performance degradation. It is one of the top reasons where model accuracy degrades over time, thus monitoring data drift helps detect model performance issues.
B: A confusion matrix is used to describe the performance of a classification model. Each row displays the instances of the true, or actual class in your dataset, and each column represents the instances of the class that was predicted by the model.
C: Hyperparameters are adjustable parameters you choose for model training that guide the training process. The HyperDrive package helps you automate choosing these parameters.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-machine-learning-interpretability-automl
Question #2*Topic 5*

HOTSPOT -
You write code to retrieve an experiment that is run from your Azure Machine Learning workspace. The run used the model interpretation support in Azure Machine Learning to generate and upload a model explanation.
Business managers in your organization want to see the importance of the features in the model. You need to print out the model features and their relative importance in an output that looks similar to the following.

Feature	Importance
0	1.5627435610083558
2	0.6077689312583112
4	0.5574002432900718
3	0.42858759955671777
1	0.3501361539771977

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

Hot Area:

Answer Area

```
# Assume required modules are imported

ws = Workspace.from_config()
feature_importances = explanation.[ from_run | list_model_explanations | from_run_id | download_model_explanation ]( workspace = ws,
                                   experiment_name='train_and_explain',
                                   run_id='train_and_explain_12345')
explanation = client.[ upload_model_explanation | list_model_explanations | run | download_model_explanation ]()

feature_importances = explanation.[ explanation | explanation_client | get_feature_importance | download_model_explanation ]()

for key, value in feature_importances.items():
        print(key, "\t", value)
```

Correct Answer:

Answer Area

```
# Assume required modules are imported

ws = Workspace.from_config()
feature_importances = explanation.from_run_id( workspace = ws,
                                   experiment_name='train_and_explain',
                                   run_id='train_and_explain_12345')
explanation = client.list_model_explanations()

feature_importances = explanation.explanation()

for key, value in feature_importances.items():
        print(key, "\t", value)
```

Box 1: from_run_id -

from_run_id(workspace, experiment_name, run_id)

Create the client with factory method given a run ID.

Returns an instance of the ExplanationClient.

Parameters -

- Workspace Workspace - An object that represents a workspace.
- experiment_name str - The name of an experiment.

➾ run_id str - A GUID that represents a run.

Box 2: list_model_explanations -
list_model_explanations returns a dictionary of metadata for all model explanations available.

Returns -
A dictionary of explanation metadata such as id, data type, explanation method, model type, and upload time, sorted by upload time

Box 3: explanation -
Reference:
https://docs.microsoft.com/en-us/python/api/azureml-contrib-interpret/azureml.contrib.interpret.explanation.explanation_client.explanationclient?view=azure-ml-py
Question #3*Topic 5*

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 train a classification model by using a logistic regression algorithm.
You must be able to explain the model's predictions by calculating the importance of each feature, both as an overall global relative importance value and as a measure of local importance for a specific set of predictions.
You need to create an explainer that you can use to retrieve the required global and local feature importance values.
Solution: Create a MimicExplainer.
Does the solution meet the goal?

- A. Yes
- B. No

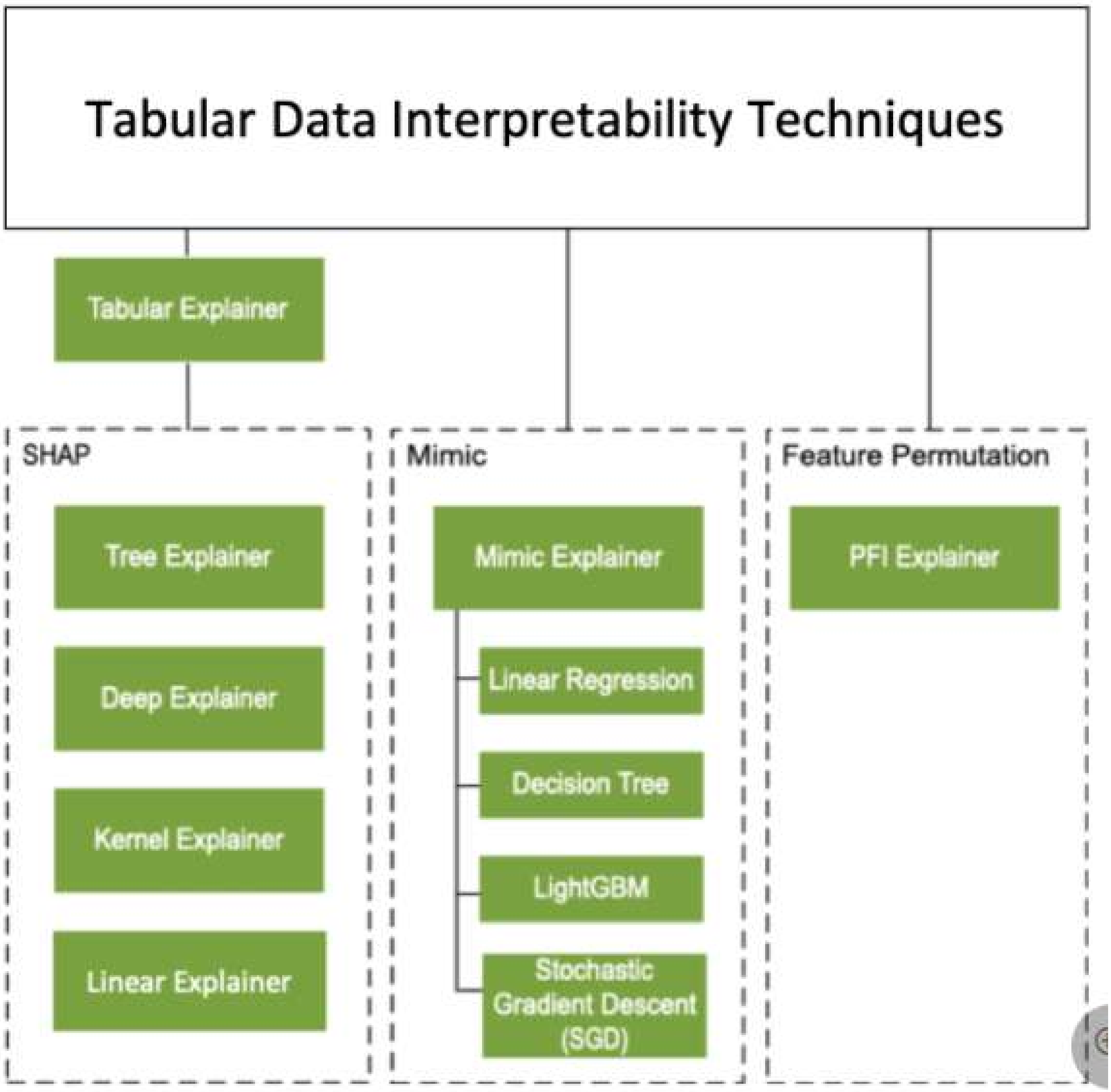
Correct Answer: *B*
Instead use Permutation Feature Importance Explainer (PFI).
Note 1: Mimic explainer is based on the idea of training global surrogate models to mimic blackbox models. A global surrogate model is an intrinsically interpretable model that is trained to approximate the predictions of any black box model as accurately as possible. Data scientists can interpret the surrogate model to draw conclusions about the black box model.
Note 2: Permutation Feature Importance Explainer (PFI): Permutation Feature Importance is a technique used to explain classification and regression models. At a high level, the way it works is by randomly shuffling data one feature at a time for the entire dataset and calculating how much the performance metric of interest changes. The larger the change, the more important that feature is.
PFI can explain the overall behavior of any underlying model but does not explain individual predictions.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-machine-learning-interpretability
Question #4*Topic 5*

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 train a classification model by using a logistic regression algorithm.
You must be able to explain the model's predictions by calculating the importance of each feature, both as an overall global relative importance value and as a measure of local importance for a specific set of predictions.
You need to create an explainer that you can use to retrieve the required global and local feature importance values.
Solution: Create a TabularExplainer.
Does the solution meet the goal?

- A. Yes
- B. No

Correct Answer: *B*
Instead use Permutation Feature Importance Explainer (PFI).
Note 1:



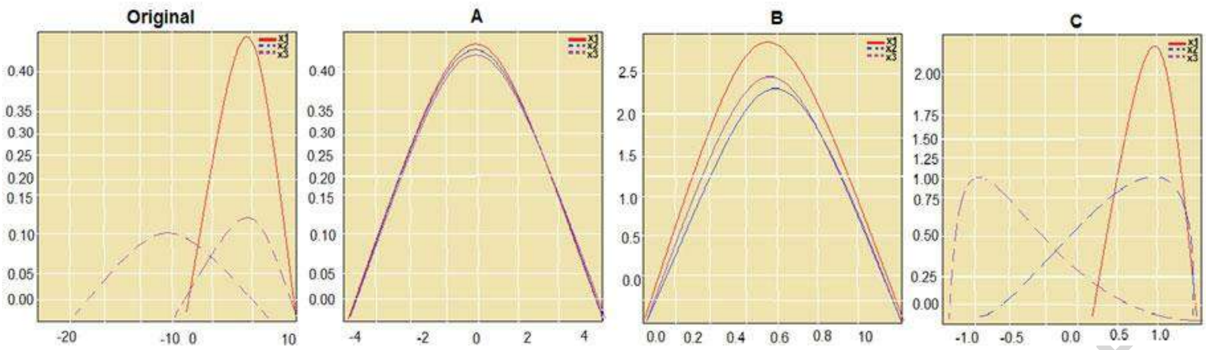
Note 2: Permutation Feature Importance Explainer (PFI): Permutation Feature Importance is a technique used to explain classification and regression models. At a high level, the way it works is by randomly shuffling data one feature at a time for the entire dataset and calculating how much the performance metric of interest changes. The larger the change, the more important that feature is. PFI can explain the overall behavior of any underlying model but does not explain individual predictions.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-machine-learning-interpretability
Question #5*Topic 5*

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 train a classification model by using a logistic regression algorithm.
You must be able to explain the model's predictions by calculating the importance of each feature, both as an overall global relative importance value and as a measure of local importance for a specific set of predictions.
You need to create an explainer that you can use to retrieve the required global and local feature importance values.
Solution: Create a PFIExplainer.
Does the solution meet the goal?

- A. Yes
- B. No

Correct Answer: *A*
Permutation Feature Importance Explainer (PFI): Permutation Feature Importance is a technique used to explain classification and regression models. At a high level, the way it works is by randomly shuffling data one feature at a time for the entire dataset and calculating how much the performance metric of interest changes. The larger the change, the more important that feature is. PFI can explain the overall behavior of any underlying model but does not explain individual predictions.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-machine-learning-interpretability
Question #6*Topic 5*

HOTSPOT -
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.

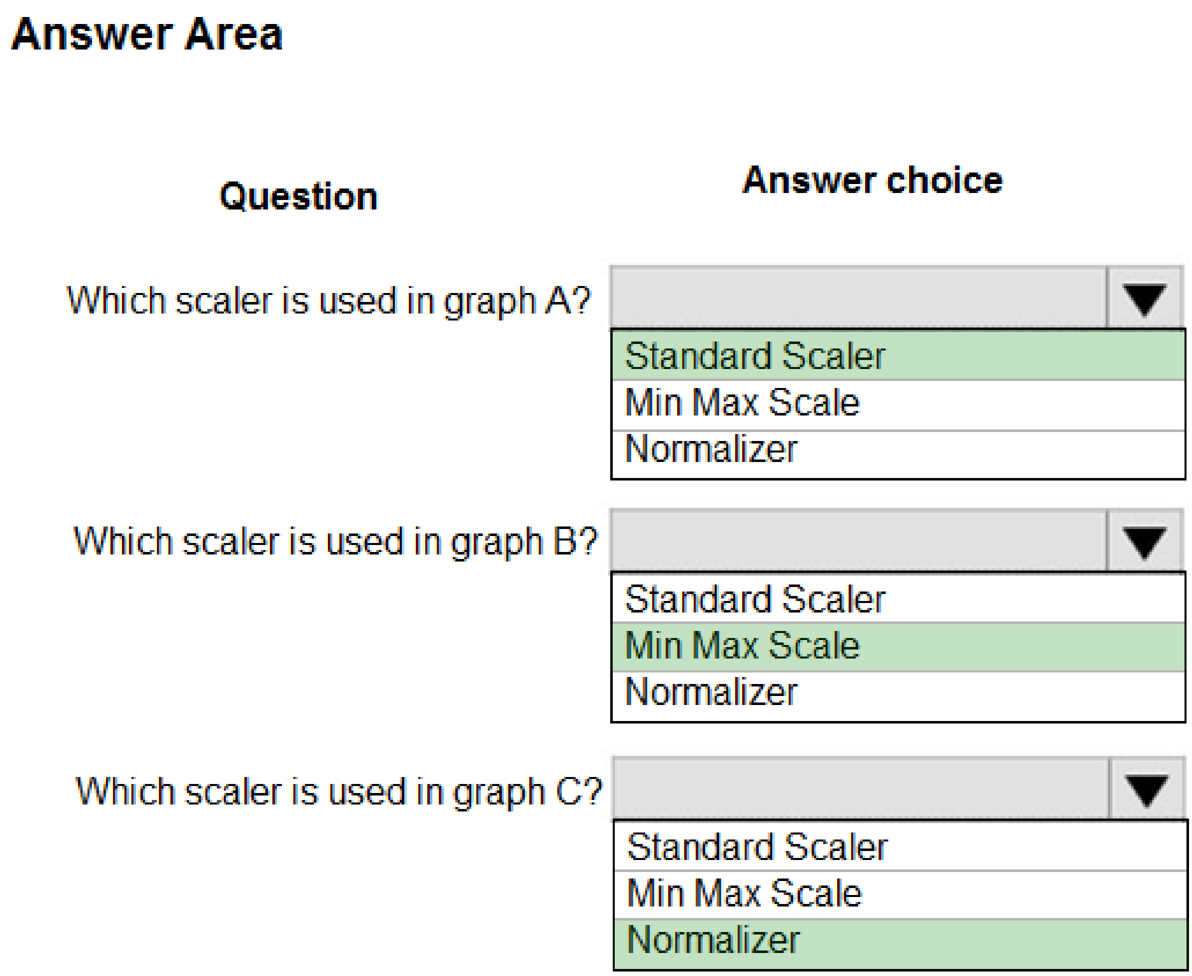
NOTE: Each correct selection is worth one point.

Hot Area:

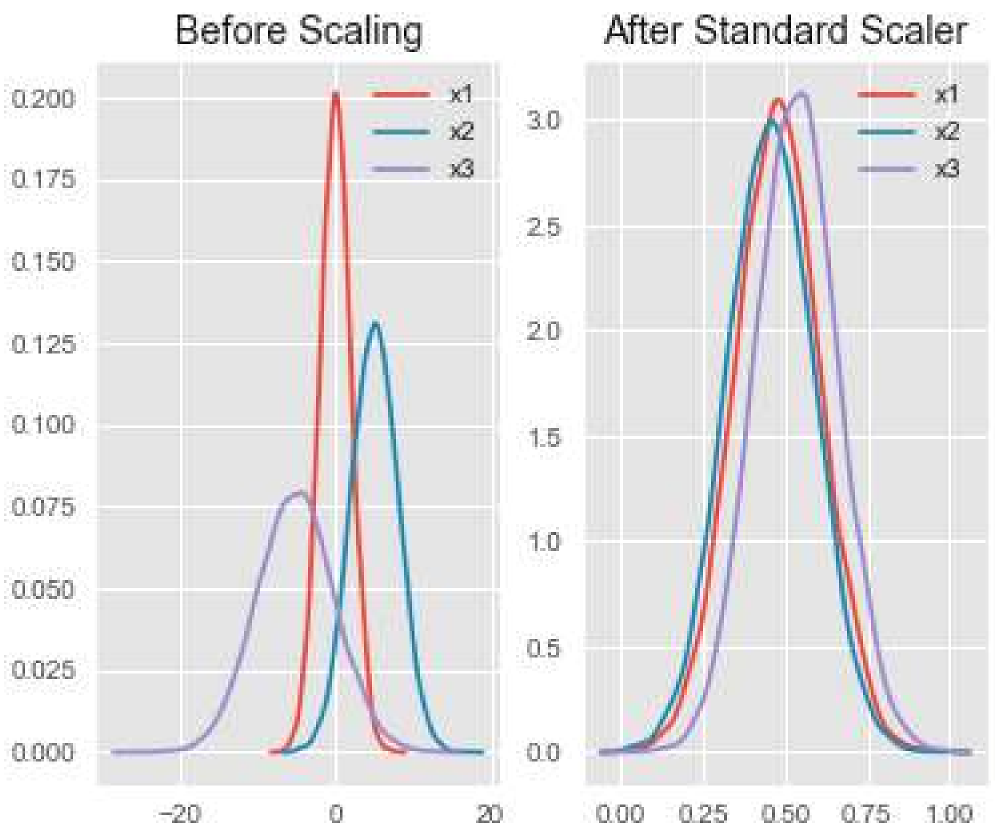
Answer Area

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

Correct Answer:

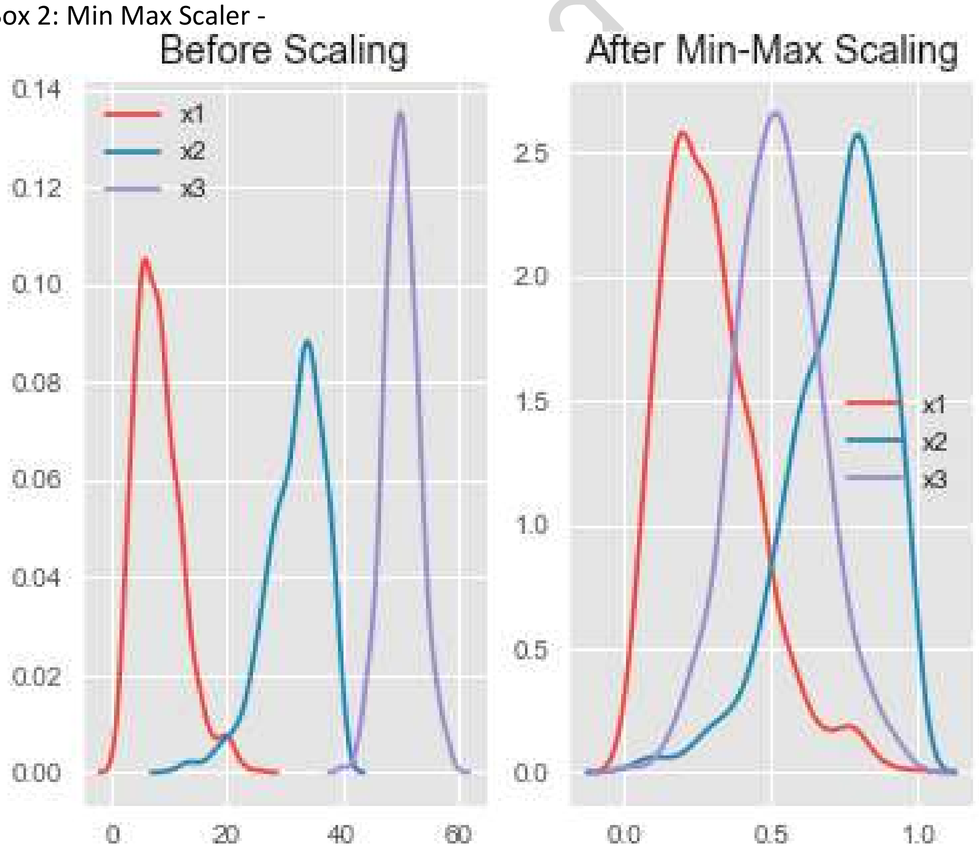


Box 1: StandardScaler -
The StandardScaler assumes your data is normally distributed within each feature and will scale them such that the distribution is now centred around 0, with a standard deviation of 1.
Example:



All features are now on the same scale relative to one another.

Box 2: Min Max Scaler -



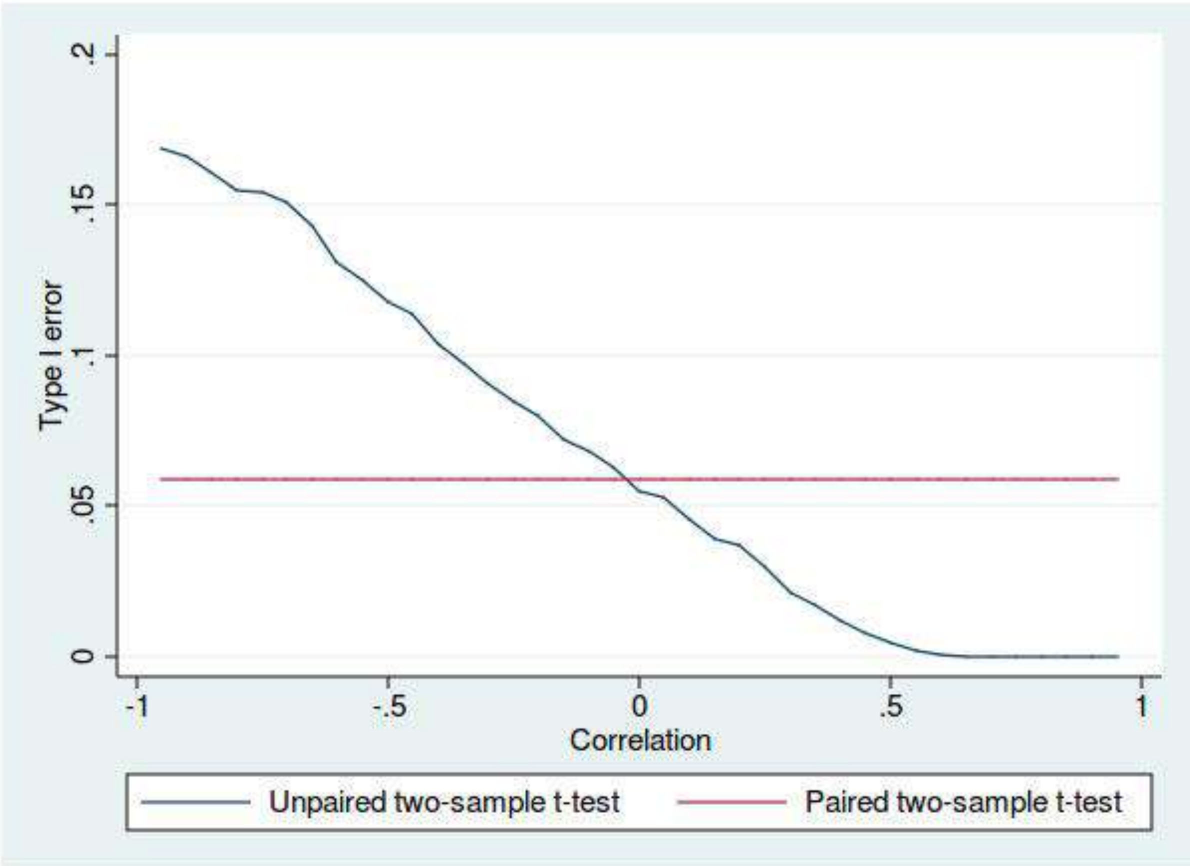
Notice that the skewness of the distribution is maintained but the 3 distributions are brought into the same scale so that they overlap.

Box 3: Normalizer -
Reference:
http://benalexkeen.com/feature-scaling-with-scikit-learn/
Question #7*Topic 5*

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. Unpaired t-test with a two-tail option
- B. Unpaired t-test with a one-tail option
- C. Paired t-test with a one-tail option
- D. Paired t-test with a two-tail option

Correct Answer: *D*
Choose a one-tail or two-tail test. The default is a two-tailed test. This is the most common type of test, in which the expected distribution is symmetric around zero.
Example: Type I error of unpaired and paired two-sample t-tests as a function of the correlation. The simulated random numbers originate from a bivariate normal distribution with a variance of 1.



Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/test-hypothesis-using-t-test https://en.wikipedia.org/wiki/Student%27s_t-test

Question #8*Topic 5*

DRAG DROP -
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 effective 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.
Select and Place:

Action	Answer area
Evaluate the probability function	
Remove duplicate rows	
Use the Filter Based Feature Selection module	
Test the hypothesis using t-Test	
Compute linear correlation	
Build a counting transform	

Correct Answer:

Action	Answer area
Evaluate the probability function	Use the Filter Based Feature Selection module
Remove duplicate rows	Build a counting transform
Use the Filter Based Feature Selection module	Test the hypothesis using t-Test
Test the hypothesis using t-Test	
Compute linear correlation	
Build a counting transform	

Step 1: Use the Filter Based Feature Selection module
Filter Based Feature Selection identifies the features in a dataset with the greatest predictive power. The module outputs a dataset that contains the best feature columns, as ranked by predictive power. It also outputs the names of the features and their scores from the selected metric.
Step 2: Build a counting transform
A counting transform creates a transformation that turns count tables into features, so that you can

apply the transformation to multiple datasets.
Step 3: Test the hypothesis using t-Test
Reference:
https://docs.microsoft.com/bs-latn-ba/azure/machine-learning/studio-module-reference/filter-based-feature-selection https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/build-counting-transform
Question #9*Topic 5*

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. Extract N-Gram Features from Text
- B. Edit Metadata
- C. Preprocess Text
- D. Apply SQL Transformation

Correct Answer: *B*
Typical metadata changes might include marking columns as features.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/edit-metadata
Question #10*Topic 5*

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

Correct Answer: *A*
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.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/evaluate-model
Question #11*Topic 5*

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

Correct Answer: *B*
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.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/evaluate-model
Question #12*Topic 5*

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

Correct Answer: *B*
Relative Squared Error, Coefficient of Determination are good metrics to evaluate the linear regression model, but the others are metrics for classification models.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/evaluate-model
Question #13*Topic 5*

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. Root Mean Square Error
- B. Coefficient of determination
- C. Recall
- D. Precision
- E. Mean absolute error

Correct Answer: *B*
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.
Incorrect Answers:
A: 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.
C: Recall is the fraction of all correct results returned by the model.
D: Precision is the proportion of true results over all positive results.
E: Mean absolute error (MAE) measures how close the predictions are to the actual outcomes; thus, a lower score is better.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/evaluate-model
Question #14*Topic 5*

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
- E. Accuracy
- F. Root Mean Square Error

Correct Answer: *B*
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.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio/evaluate-model-performance#evaluating-a-binary-classification-model
Question #15*Topic 5*

HOTSPOT -
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.
Hot Area:

Answer Area

Which algorithm minimizes differences between actual and predicted values?

- Bayesian Linear Regression
- Neutral 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 False.
- Set the Create trainer mode option to Parameter Range.
- Increase the number of epochs.
- Decrease the number of epochs.

Correct Answer:

Answer Area

Which algorithm minimizes differences between actual and predicted values?

- Bayesian Linear Regression
- Neutral Network Regression
- Boosted Decision Tree Regression
- Linear Regression
- Decision Forest Regression

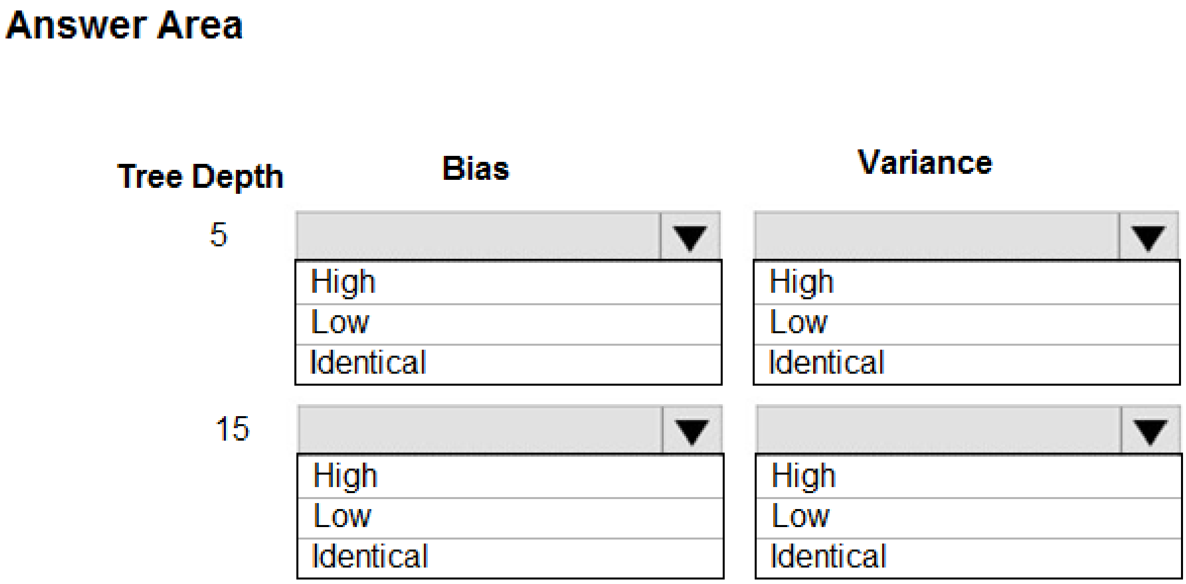
Which approach should you use to find the best parameters for a Linear Regression model for the Online Gradient Descent method?

- Set the Decrease learning rate option to True.
- Set the Decrease learning rate option to False.
- Set the Create trainer mode option to Parameter Range.
- Increase the number of epochs.
- Decrease the number of epochs.

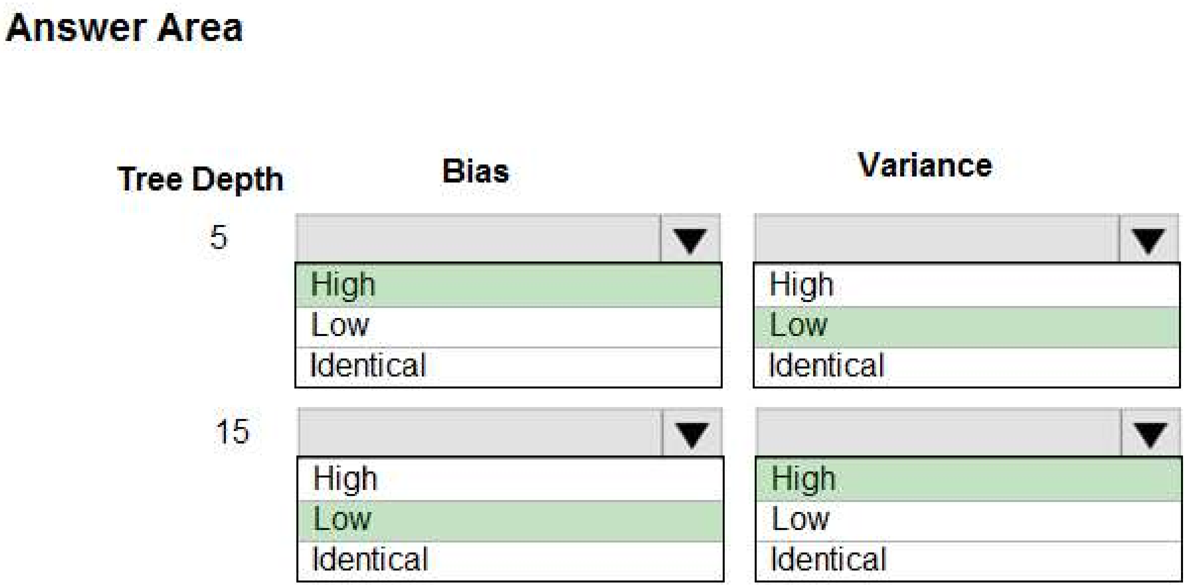
Box 1: Boosted Decision Tree Regression
Mean absolute error (MAE) measures how close the predictions are to the actual outcomes; thus, a lower score is better.
Box 2:
Online Gradient Descent: If you want the algorithm to find the best parameters for you, set Create trainer mode option to Parameter Range. You can then specify multiple values for the algorithm to try.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/evaluate-model https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/linear-regression
Question #16*Topic 5*

HOTSPOT -
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.

Hot Area:



Correct Answer:



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 biasג€"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.
Reference:
https://machinelearningmastery.com/gentle-introduction-to-the-bias-variance-trade-off-in-machine-learning/
Question #17*Topic 5*

DRAG DROP -
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.
Select and Place:

Answer Area

Modules

- Two-Class Boosted Decision Tree
- Partition and Sample
- Tune Model Hyperparameters
- Split Data

Step	Module
Define the parameter scope	
Define the cross-validation settings	
Define the metric	
Train, evaluate, and compare	

Correct Answer:

Answer Area

Modules

- Two-Class Boosted Decision Tree
- Partition and Sample
- Tune Model Hyperparameters
- Split Data

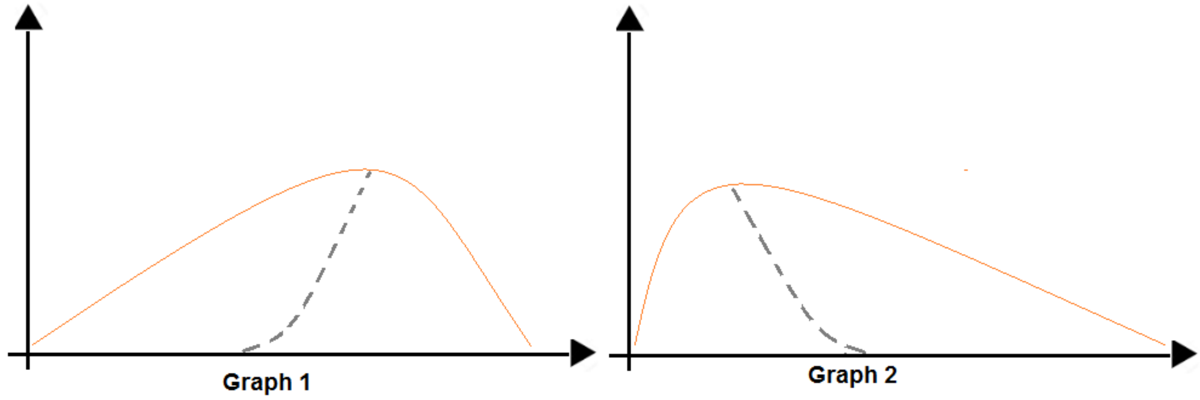
Step	Module
Define the parameter scope	Split Data
Define the cross-validation settings	Partition and Sample
Define the metric	Two-Class Boosted Decision Tree
Train, evaluate, and compare	Tune Model Hyperparameters

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.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/partition-and-sample

Question #18*Topic 5*

HOTSPOT -
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.
Hot Area:

Answer Area

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

Correct Answer:

Answer Area

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

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.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/compute-elementary-statistics
Question #19*Topic 5*

You are a data scientist building a deep convolutional neural network (CNN) for image classification.
The CNN model you build 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. Add an additional dense layer with 512 input units.
- B. Add L1/L2 regularization.
- C. Use training data augmentation.
- D. Reduce the amount of training data.
- E. Add an additional dense layer with 64 input units.

Correct Answer: *BD*
B: Weight regularization provides an approach to reduce the overfitting of a deep learning neural network model on the training data and improve the performance of the model on new data, such as the holdout test set.
Keras provides a weight regularization API that allows you to add a penalty for weight size to the loss function.
Three different regularizer instances are provided; they are:
⇨ L1: Sum of the absolute weights.
⇨ L2: Sum of the squared weights.
⇨ L1L2: Sum of the absolute and the squared weights.
D: Because a fully connected layer occupies most of the parameters, it is prone to overfitting. One method to reduce overfitting is dropout. At each training stage, individual nodes are either "dropped out" of the net with probability 1-p or kept with probability p, so that a reduced network is left;

incoming and outgoing edges to a dropped-out node are also removed.
By avoiding training all nodes on all training data, dropout decreases overfitting.
Reference:
https://machinelearningmastery.com/how-to-reduce-overfitting-in-deep-learning-with-weight-regularization/ https://en.wikipedia.org/wiki/Convolutional_neural_network
Question #20*Topic 5*

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, Accuracy, Precision, Recall, F1 score, and AUC.
Does the solution meet the goal?

- A. Yes
- B. No

Correct Answer: *B*
Accuracy, Precision, Recall, F1 score, and AUC are metrics for evaluating classification models.
Note: Mean Absolute Error, Root Mean Absolute Error, Relative Absolute Error are OK for the linear regression model.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/evaluate-model
Question #21*Topic 5*

You are building a binary classification model by using a supplied training set.
The training set is imbalanced between two classes.
You need to resolve the data imbalance.
What are three possible ways to achieve this goal? Each correct answer presents a complete solution.
NOTE: Each correct selection is worth one point.

- A. Penalize the classification
- B. Resample the dataset using undersampling or oversampling
- C. Normalize the training feature set
- D. Generate synthetic samples in the minority class
- E. Use accuracy as the evaluation metric of the model

Correct Answer: *ABD*
A: Try Penalized Models -
You can use the same algorithms but give them a different perspective on the problem.
Penalized classification imposes an additional cost on the model for making classification mistakes on the minority class during training. These penalties can bias the model to pay more attention to the minority class.

B: You can change the dataset that you use to build your predictive model to have more balanced data.
This change is called sampling your dataset and there are two main methods that you can use to even-up the classes:
➾ Consider testing under-sampling when you have an a lot data (tens- or hundreds of thousands of instances or more)
➾ Consider testing over-sampling when you don't have a lot of data (tens of thousands of records or less)
D: Try Generate Synthetic Samples
A simple way to generate synthetic samples is to randomly sample the attributes from instances in the minority class.
Reference:
https://machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset/
Question #22*Topic 5*

HOTSPOT -
You train a classification model by using a decision tree algorithm.
You create an estimator by running the following Python code. The variable feature_names is a list of all feature names, and class_names is a list of all class names. from interpret.ext.blackbox import TabularExplainer explainer = TabularExplainer(model, x_train, features=feature_names, classes=class_names)
You need to explain the predictions made by the model for all classes by determining the importance of all features.
For each of the following statements, select Yes if the statement is true. Otherwise, select No.
NOTE: Each correct selection is worth one point.
Hot Area:

Answer Area

	Yes	No
The SHAP TreeExplainer will be used to interpret the model.	○	○
If you omit the features and classes parameters in the TabularExplainer instantiation, the explainer still works as expected.	○	○
You could interpret the model by using a MimicExplainer instead of a TabularExplainer.	○	○

Correct Answer:

Answer Area

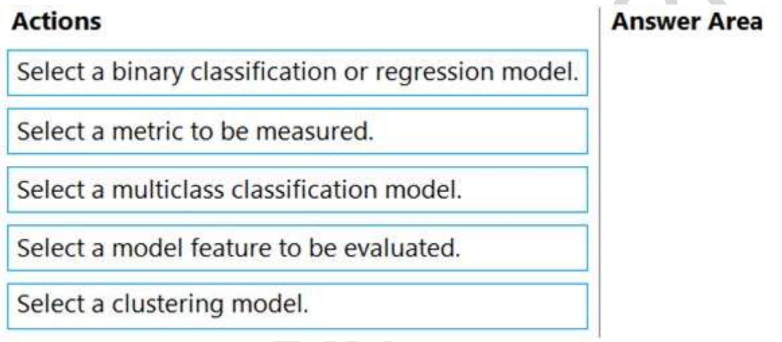
	Yes	No
The SHAP TreeExplainer will be used to interpret the model.	**○**	○
If you omit the features and classes parameters in the TabularExplainer instantiation, the explainer still works as expected.	**○**	○
You could interpret the model by using a MimicExplainer instead of a TabularExplainer.	○	**○**

Box 1: Yes -
TabularExplainer calls one of the three SHAP explainers underneath (TreeExplainer, DeepExplainer, or KernelExplainer).

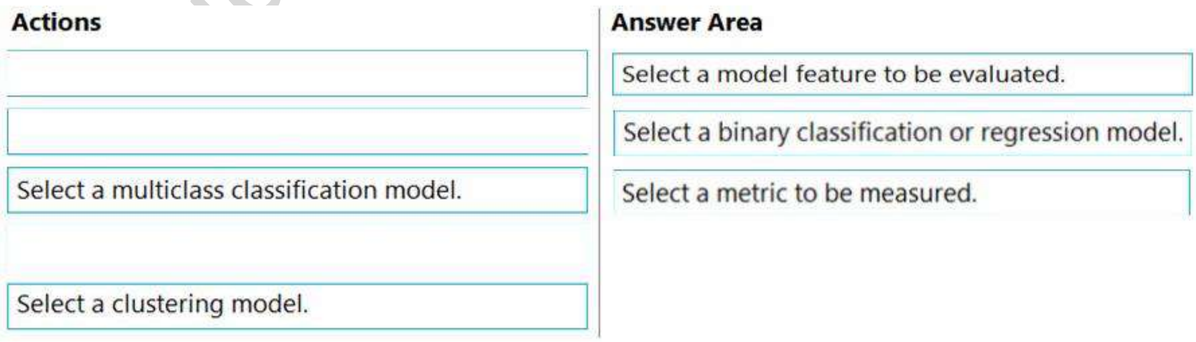
Box 2: Yes -
To make your explanations and visualizations more informative, you can choose to pass in feature names and output class names if doing classification.

Box 3: No -
TabularExplainer automatically selects the most appropriate one for your use case, but you can call each of its three underlying explainers underneath
(TreeExplainer, DeepExplainer, or KernelExplainer) directly.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-machine-learning-interpretability-aml
Question #23*Topic 5*

DRAG DROP -
You have several machine learning models registered in an Azure Machine Learning workspace.
You must use the Fairlearn dashboard to assess fairness in a selected model.
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.
Select and Place:



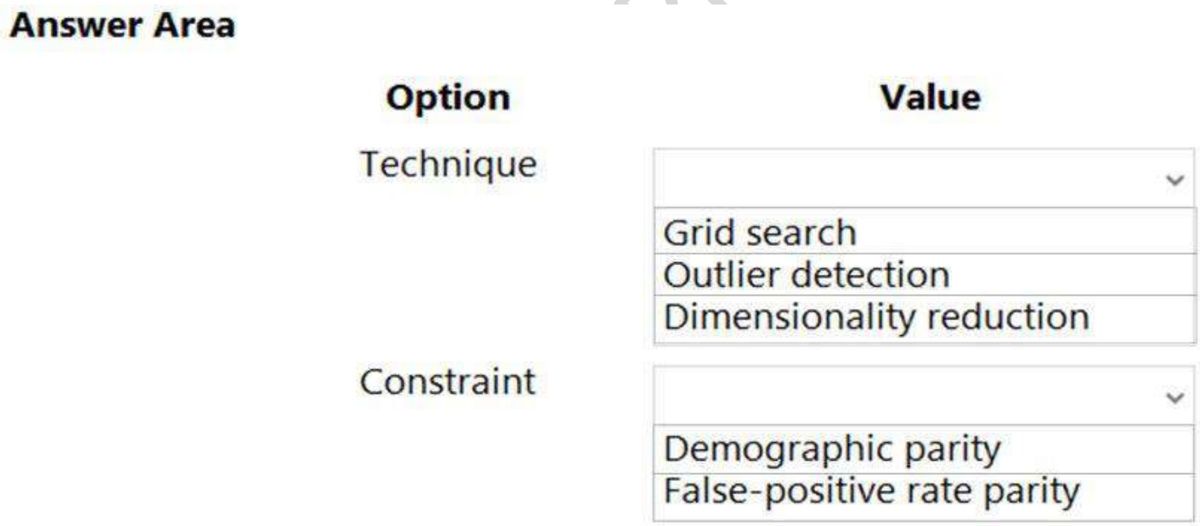
Correct Answer:



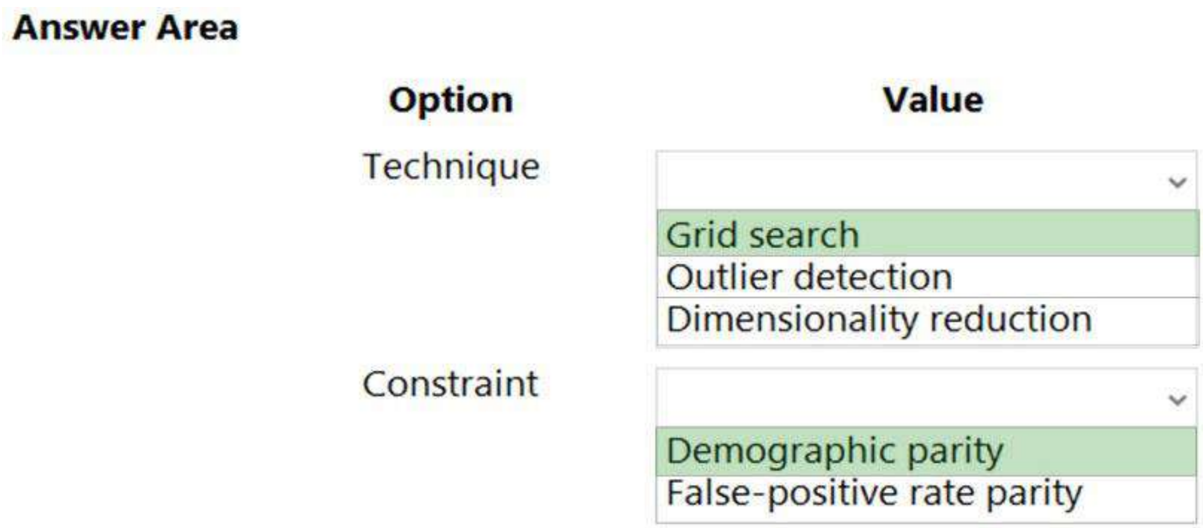
Step 1: Select a model feature to be evaluated.
Step 2: Select a binary classification or regression model.
Register your models within Azure Machine Learning. For convenience, store the results in a dictionary, which maps the id of the registered model (a string in name:version format) to the predictor itself.

Example:
model_dict = {}
lr_reg_id = register_model("fairness_logistic_regression", lr_predictor) model_dict[lr_reg_id] = lr_predictor svm_reg_id = register_model("fairness_svm", svm_predictor) model_dict[svm_reg_id] = svm_predictor
Step 3: Select a metric to be measured
Precompute fairness metrics.
Create a dashboard dictionary using Fairlearn's metrics package.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-machine-learning-fairness-aml
Question #24*Topic 5*

HOTSPOT -
A biomedical research company plans to enroll people in an experimental medical treatment trial. You create and train a binary classification model to support selection and admission of patients to the trial. The model includes the following features: Age,
Gender, and Ethnicity.
The model returns different performance metrics for people from different ethnic groups.
You need to use Fairlearn to mitigate and minimize disparities for each category in the Ethnicity feature.
Which technique and constraint should you use? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.
Hot Area:



Correct Answer:



Box 1: Grid Search -
Fairlearn open-source package provides postprocessing and reduction unfairness mitigation algorithms: ExponentiatedGradient, GridSearch, and ThresholdOptimizer.
Note: The Fairlearn open-source package provides postprocessing and reduction unfairness mitigation algorithms types:
- Reduction: These algorithms take a standard black-box machine learning estimator (e.g., a LightGBM model) and generate a set of retrained models using a sequence of re-weighted training datasets.
- Post-processing: These algorithms take an existing classifier and the sensitive feature as input.

Box 2: Demographic parity -
The Fairlearn open-source package supports the following types of parity constraints: Demographic parity, Equalized odds, Equal opportunity, and Bounded group loss.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/concept-fairness-ml
Question #25*Topic 5*

You create a binary classification model. The model is registered in an Azure Machine Learning workspace. You use the Azure Machine Learning Fairness SDK to assess the model fairness.
You develop a training script for the model on a local machine.
You need to load the model fairness metrics into Azure Machine Learning studio.
What should you do?

- A. Implement the download_dashboard_by_upload_id function
- B. Implement the create_group_metric_set function
- C. Implement the upload_dashboard_dictionary function
- D. Upload the training script

Correct Answer: *C*
import azureml.contrib.fairness package to perform the upload: from azureml.contrib.fairness import upload_dashboard_dictionary, download_dashboard_by_upload_id
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/how-to-machine-learning-fairness-aml
Question #26*Topic 5*

You have a dataset that includes confidential data. You use the dataset to train a model.
You must use a differential privacy parameter to keep the data of individuals safe and private.
You need to reduce the effect of user data on aggregated results.
What should you do?

- A. Decrease the value of the epsilon parameter to reduce the amount of noise added to the data
- B. Increase the value of the epsilon parameter to decrease privacy and increase accuracy
- C. Decrease the value of the epsilon parameter to increase privacy and reduce accuracy
- D. Set the value of the epsilon parameter to 1 to ensure maximum privacy

Correct Answer: *C*
Differential privacy tries to protect against the possibility that a user can produce an indefinite number of reports to eventually reveal sensitive data. A value known as epsilon measures how noisy, or private, a report is. Epsilon has an inverse relationship to noise or privacy. The lower the epsilon, the more noisy (and private) the data is.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/concept-differential-privacy
Question #27*Topic 5*

HOTSPOT -
You create an Azure Machine Learning workspace and load a Python training script named train.py in the src subfolder. The dataset used to train your model is available locally.
You run the following script to train the model:

```
ws = Workspace.from_config()
experiment = Experiment(workspace=ws, name='nlp-experiment-train')

cpu_cluster_name = "cpu-cluster"
try:
    cpu_cluster = ComputeTarget(workspace=ws, name=cpu_cluster_name)
except ComputeTargetException:
    compute_config = AmlCompute.provisioning_configuration(vm_size='STANDARD_D2_V2', max_nodes=4)
    cpu_cluster = ComputeTarget.create(ws, cpu_cluster_name, compute_config)

cpu_cluster.wait_for_completion(show_output=True)

config = ScriptRunConfig(source_directory='./src',
                         script='train.py',
                         compute_target='cpu-cluster')

env = Environment.from_conda_specification(
    name='pytorch-env',
    file_path='./.azureml/pytorch-env.yml'
)
config.run_config.environment = env

run = experiment.submit(config)
```

Instructions: For each of the following statements, select Yes if the statement is true. Otherwise, select No.
NOTE: Each correct selection is worth one point.

Hot Area:

Statements	Yes	No
The script will use local compute resources and a new Azure Machine Learning compute will be created upon failure.	○	○
The dataset used during the training phase is automatically loaded in a new datastore.	○	○
A new environment object is created from a local Conda environment.	○	○

Correct Answer:

Statements	Yes	No
The script will use local compute resources and a new Azure Machine Learning compute will be created upon failure.	○	○
The dataset used during the training phase is automatically loaded in a new datastore.	○	○
A new environment object is created from a local Conda environment.	○	○

Question #28*Topic 5*

You develop a machine learning project on a local machine. The project uses the Azure Machine Learning SDK for Python. You use Git as version control for scripts.
You submit a training run that returns a Run object.
You need to retrieve the active Git branch for the training run.
Which two code segments should you use? Each correct answer presents part of the solution.
NOTE: Each correct selection is worth one point.

- A. details = run.get_environment()
- B. details.properties['azureml.git.branch']
- C. details.properties['azureml.git.commit']
- D. details = run.get_details()

Correct Answer: *BC*

Question #29*Topic 5*

You are attaching an Azure Databricks-based compute resource to an Azure Machine Learning development workspace.
You need to configure parameters to attach the resource.
Which three parameters should you use? Each correct answer presents part of the solution.
NOTE: Each correct selection is worth one point.

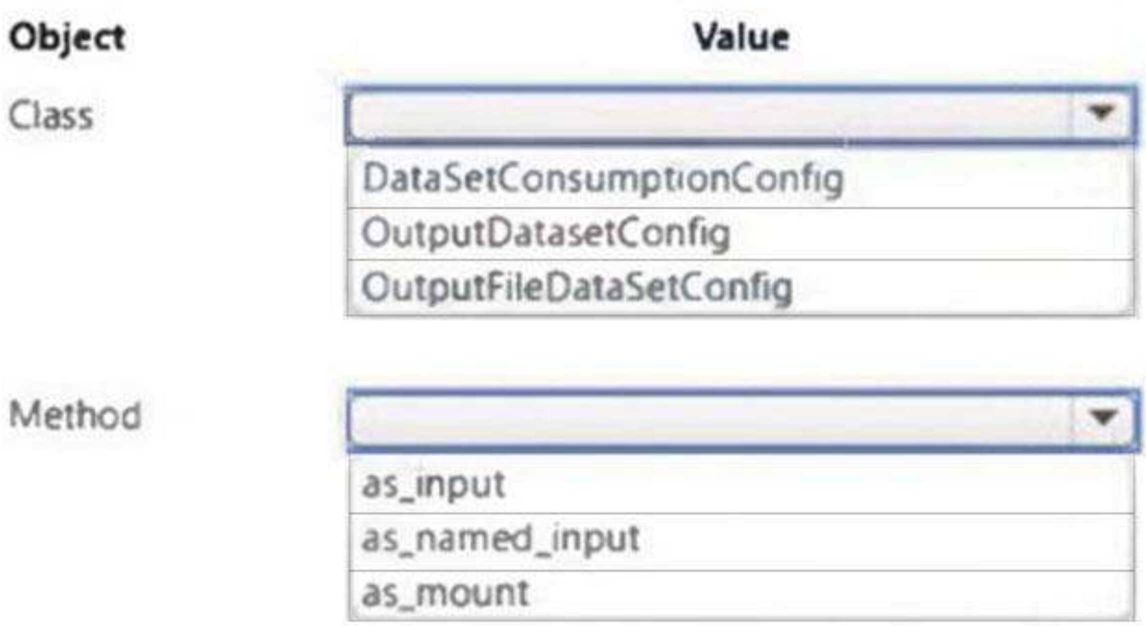
- A. Workspace name
- B. Compute name
- C. Workspace user credentials
- D. Workspace resource ID

- E. Access token

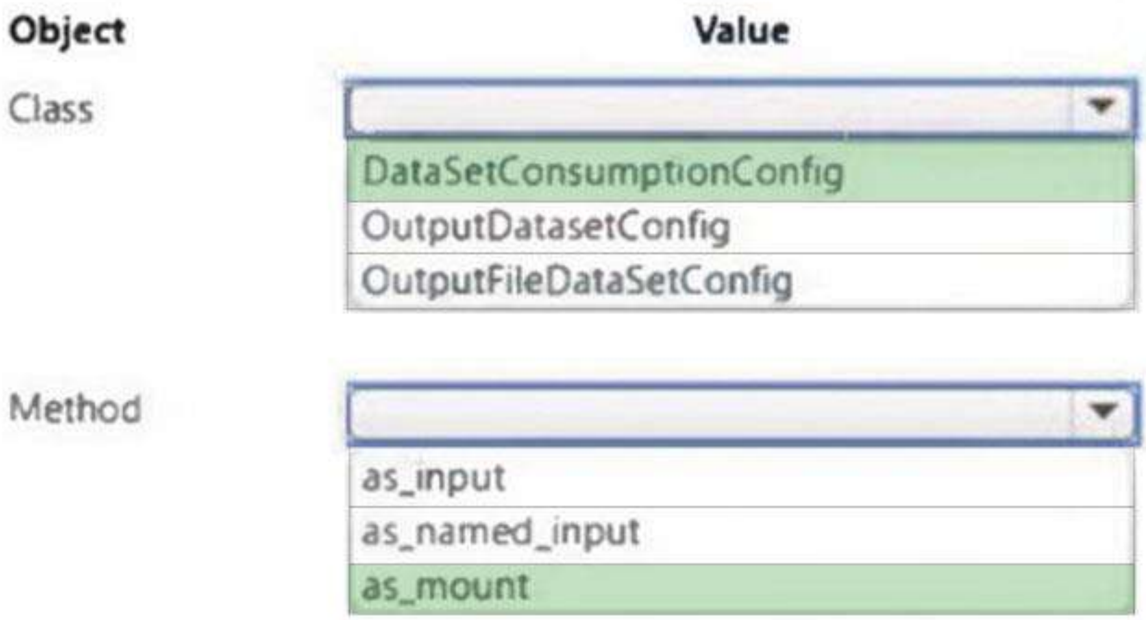
Correct Answer: *ABE*

Question #30*Topic 5*

HOTSPOT -
You are developing a two-step Azure Machine Learning pipeline by using the Azure Machine Learning SDK for Python.
The pipeline must pass temporary data from the first step to the second step.
You need to configure the second step to ensure that it can use the temporary data from the first step.
Which class and method should you use? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.
Hot Area:

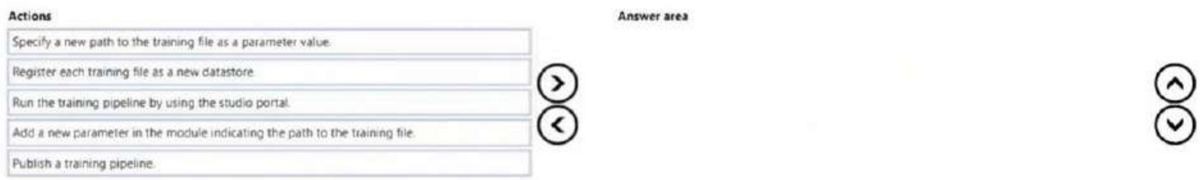


Correct Answer:



Question #31*Topic 5*

DRAG DROP -
You use a training pipeline in the Azure Machine Learning designer. You register a datastore named ds1. The datastore contains multiple training data files. You use the Import Data module with the configured datastore.
You need to retrain a model on a different set of data files.
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.
Select and Place:



Correct Answer:

Question #32*Topic 5*

You create a binary classification model. You use the Fairlearn package to assess model fairness.
You must eliminate the need to retrain the model.
You need to implement the Fairlearn package.
Which algorithm should you use?

- A. fairlearn.reductions.ExponentiatedGradient
- B. fairlearn.postprocessing.ThresholdOptimizer
- C. fairlearnpreprocessing.CorrelationRemover
- D. fairlearn.reductions.GridSearch

Correct Answer: *C*

Question #33*Topic 5*

You have an Azure Machine Learning workspace named workspace1.
You must add a datastore that connects an Azure Blob storage container to workspace1. You must be able to configure a privilege level.
You need to configure authentication.
Which authentication method should you use?

- A. Service principal
- B. Account key
- C. SAS token
- D. Managed identity

Correct Answer: *D*

Question #34*Topic 5*

You plan to create a compute instance as part of an Azure Machine Learning development workspace.
You must interactively debug code running on the compute instance by using Visual Studio Code Remote.
You need to provision the compute instance.
What should you do?

- A. Enable Remote Desktop Protocol (RDP) access.
- B. Modify role-based access control (RBAC) settings at the workspace level.
- C. Enable Secure Shell Protocol (SSH) access.
- D. Modify role-based access control (RBAC) settings at the compute instance level.

Correct Answer: *B*

Question #35*Topic 5*

You have a dataset that contains salary information for users. You plan to generate an aggregate salary report that shows average salaries by city.
Privacy of individuals must be preserved without impacting accuracy, completeness, or reliability of the data. The aggregation must be statistically consistent with the distribution of the original data.
You must return an approximation of the data instead of the raw data.
You need to apply a differential privacy approach.
What should you do?

- A. Add noise to the salary data during the analysis
- B. Encrypt the salary data before analysis
- C. Remove the salary data

- D. Convert the salary data to the average column value

Correct Answer: *D*

Question #36*Topic 5*

HOTSPOT
-

You have a binary classifier that predicts positive cases of diabetes within two separate age groups.

The classifier exhibits a high degree of disparity between the age groups.

You need to modify the output of the classifier to maximize its degree of fairness across the age groups and meet the following requirements:

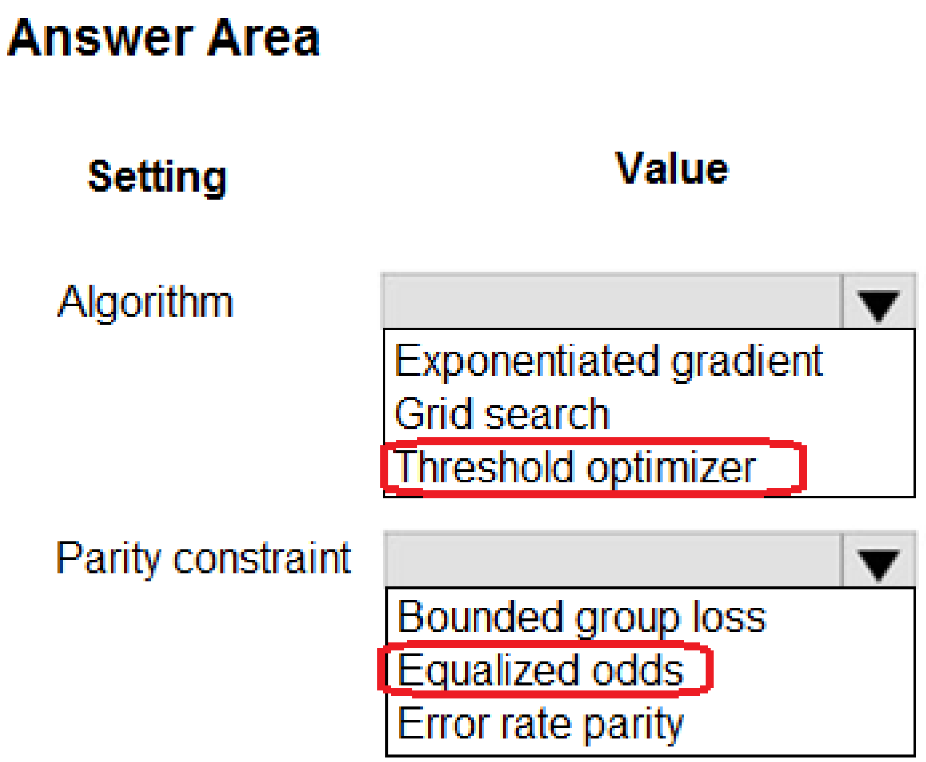
• Eliminate the need to retrain the model on which the classifier is based.
• Minimize the disparity between true positive rates and false positive rates across age groups.

Which algorithm and parity constraint should you use? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Answer Area

Setting	Value
Algorithm	Exponentiated gradient Grid search Threshold optimizer
Parity constraint	Bounded group loss Equalized odds Error rate parity



Correct Answer:

Question #37*Topic 5*

You create an Azure Machine Learning workspace. You train an MLflow-formatted regression model by using tabular structured data.

You must use a Responsible AI dashboard to assess the model.

You need to use the Azure Machine Learning studio UI to generate the Responsible AI dashboard.

What should you do first?

- A. Convert the model from the MLflow format to a custom format.
- B. Register the model with the workspace.
- C. Create the model explanations.
- D. Deploy the model to a managed online endpoint.

Correct Answer: *B*

Question #: 38
Topic #: 5
[All PL-100 Questions]

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.

A company receives its marketing campaign performance report as a Microsoft Excel file. The company stores the file in Microsoft SharePoint.

A user updates the Excel file daily with updated data.

The company needs a solution that automatically displays the latest analytics.

You need to create a basic report that contains pie charts that display the most profitable channels.

You need to propose the right solution.

Solution: Microsoft Power Apps canvas app having an Excel file imported into the app and Pie chart control.

Does the solution meet the goal?

- A. Yes
- B. No

Correct Answer: *B*

Topic 6 - Question Set 6
Question #1*Topic 6*

You create an Azure Machine Learning workspace.

You must configure an event-driven workflow to automatically trigger upon completion of training runs in the workspace. The solution must minimize the administrative effort to configure the trigger.

You need to configure an Azure service to automatically trigger the workflow.

Which Azure service should you use?

- A. Event Grid subscription
- B. Azure Automation runbook
- C. Event Hubs Capture
- D. Event Hubs consumer

Correct Answer: *A*

Question #2*Topic 6*
HOTSPOT
-

You plan to implement an Azure Machine Learning solution.

You have the following requirements:

- Run a Jupyter notebook to interactively train a machine learning model.
- Deploy assets and workflows for machine learning proof of concept by using scripting rather than custom programming.

You need to select a development technique for each requirement.

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

NOTE: Each correct selection is worth one point.

Answer Area

Requirement	Development tool
Run a Jupyter notebook to interactively train a machine learning model.	Azure CLI Azure Machine Learning studio Azure Machine Learning Python SDK Azure Machine Learning REST
Deploy assets and workflows for machine learning proof of concept by using scripting rather than custom programming.	Azure CLI Azure Machine Learning studio Azure Machine Learning Python SDK Azure Machine Learning REST

Correct Answer:

Answer Area

Requirement	Development tool
Run a Jupyter notebook to interactively train a machine learning model.	Azure CLI Azure Machine Learning studio Azure Machine Learning Python SDK Azure Machine Learning REST
Deploy assets and workflows for machine learning proof of concept by using scripting rather than custom programming.	Azure CLI Azure Machine Learning studio Azure Machine Learning Python SDK Azure Machine Learning REST

Question #3

Topic 6

You have an Azure Machine Learning (ML) model deployed to an online endpoint.

You need to review container logs from the endpoint by using Azure ML Python SDK v2. The logs must include the console log from the inference server, with print/log statements from the model's scoring script.

What should you do first?

- A. Connect by using SSH to the inference server.
- B. Create an instance of the MLClient class.
- C. Connect by using Docker tools to the inference server.
- D. Create an instance of the OnlineDeploymentOperations class.

Correct Answer: *B*

Question #: 4
Topic #: 6

You use an Azure Machine Learning workspace.

You must monitor cost at the endpoint and deployment level.

You have a trained model that must be deployed as an online endpoint. Users must authenticate by using Microsoft Entra ID.

What should you do?

- A. Deploy the model to Azure Kubernetes Service (AKS). During deployment, set the token_auth_mode parameter of the target configuration object to true.
- B. Deploy the model to Azure Kubernetes Service (AKS). During deployment, set the auth_mode parameter to configure the authentication type.
- C. Deploy the model to a managed online endpoint. During deployment, set the auth_mode parameter to configure the authentication type.
- D. Deploy the model to a managed online endpoint. During deployment, set the token_auth_mode parameter of the target configuration object to true.

Correct Answer: *C*

Topic 7 - Testlet 1
Question #1*Topic 7*

Introductory Info Case study -

Overview -
You are a data scientist in a company that provides data science for professional sporting events. Models will use 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.
Assess 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 -
Media used for penalty event detection will be provided by consumer devices. Media may include images and videos captured during the sporting event and shared 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 video transcripts or radio commentary, and logs from related social media feeds captured during the sporting events.
Crowd sentiment will include audio recordings submitted by event attendees in both mono and stereo formats.

Penalty detection and sentiment -
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.

Advertisements -
During the initial weeks in production, the following was observed:
Ad response rated 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 uncorrelated features.
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 slow.
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 shows 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.

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.
Sampling must guarantee mutual and collective exclusively between 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.
Ad response models must support non-linear boundaries of features.
The ad propensity model uses a cut threshold is 0.45 and retrains occur if weighted Kappa deviated 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:

QuestionYou 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

Correct Answer: *C*
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.
Reference:
https://docs.microsoft.com/en-us/cognitive-toolkit/post-batch-normalization-statistics

Question #2*Topic 7*

Introductory Info Case study -

Overview -
You are a data scientist in a company that provides data science for professional sporting events.
Models will use 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.
Assess 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 -
Media used for penalty event detection will be provided by consumer devices. Media may include images and videos captured during the sporting event and shared 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 video transcripts or radio commentary, and logs from related social media feeds captured during the sporting events.

Crowd sentiment will include audio recordings submitted by event attendees in both mono and stereo formats.

Penalty detection and sentiment -
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.

Advertisements -
During the initial weeks in production, the following was observed:
Ad response rated 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 uncorrelated features.
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 slow.
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 shows 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.
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.
Sampling must guarantee mutual and collective exclusively between 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.
Ad response models must support non-linear boundaries of features.
The ad propensity model uses a cut threshold is 0.45 and retrains occur if weighted Kappa deviated 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:

QuestionHOTSPOT -

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.
Hot Area:

Answer Area

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

train_smapler = deeplearninglib.DistributedSampler.(penalty_video_dataset)
train_sampler = deeplearninglib.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)
..
```

Correct Answer:

Answer Area

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

train_smapler = deeplearninglib.DistributedSampler.(penalty_video_dataset)
train_sampler = deeplearninglib.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)
..
```

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)
Incorrect Answers: ..SGD..
Scenario: All penalty detection models show inference phases using a Stochastic Gradient Descent (SGD) are running too slow.
Box 4: .. nn.parallel.DistributedDataParallel..
DistributedSampler(Sampler): The sampler that restricts data loading to a subset of the dataset.
It is especially useful in conjunction with :class:`torch.nn.parallel.DistributedDataParallel`.
Reference:
https://github.com/pytorch/pytorch/blob/master/torch/utils/data/distributed.py
Question #3*Topic 7*

Introductory Info Case study -

Overview -
You are a data scientist in a company that provides data science for professional sporting events. Models will use 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.
Assess 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 -
Media used for penalty event detection will be provided by consumer devices. Media may include images and videos captured during the sporting event and shared 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 video transcripts or radio commentary, and logs from related social media feeds captured during the sporting events.
Crowd sentiment will include audio recordings submitted by event attendees in both mono and stereo formats.

Penalty detection and sentiment -
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.

Advertisements -
During the initial weeks in production, the following was observed:
Ad response rated 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 uncorrelated features.
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

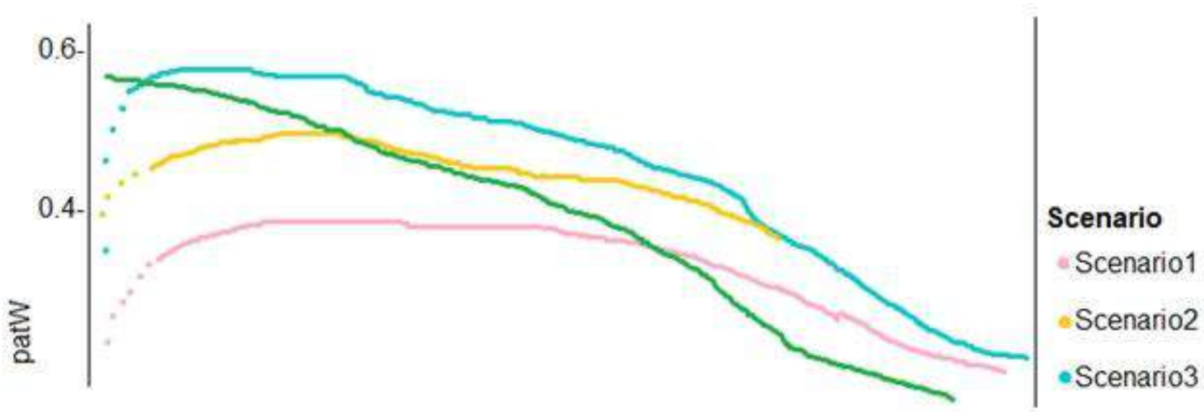
running too slow.
Audio samples show that the length of a catch phrase varies between 25%-47% depending on region
The performance of the global penalty detection models shows lower variance but higher bias when comparing training and validation sets. Before implementing any feature changes, you must confirm the bias and variance using all training and validation cases.
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.
Sampling must guarantee mutual and collective exclusively between 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.
Ad response models must support non-linear boundaries of features.
The ad propensity model uses a cut threshold is 0.45 and retrains occur if weighted Kappa deviated 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

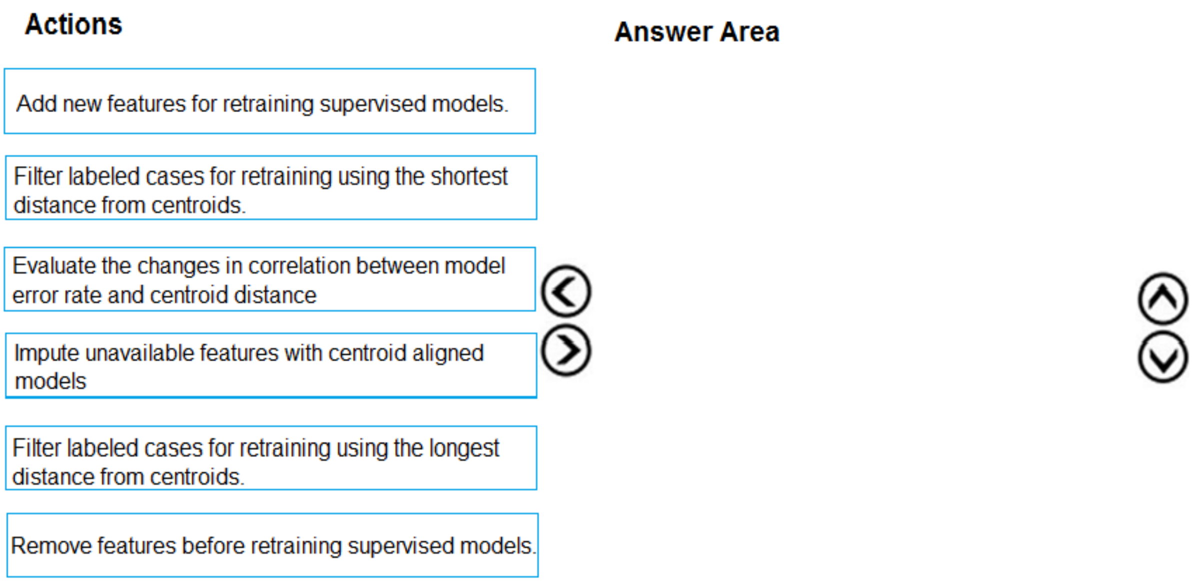
The ad propensity model uses proposed cost factors shown in the following diagram:

		Actual	
		1	0
Predicted	0	1	5
	1	5	1

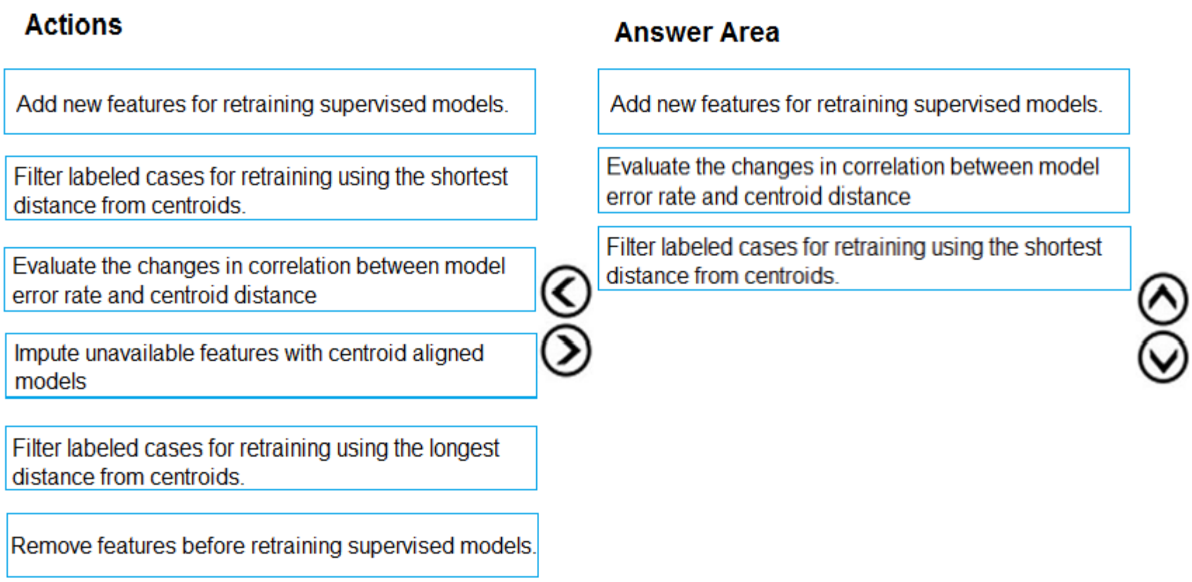
Performance curves of current and proposed cost factor scenarios are shown in the following diagram:



QuestionDRAG DROP -
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.
Select and Place:



Correct Answer:



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.
Reference:
https://en.wikipedia.org/wiki/Nearest_centroid_classifier
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/sweep-clustering

Question #4*Topic 7*

Introductory Info Case study -

Overview -
You are a data scientist in a company that provides data science for professional sporting events.
Models will use 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.
Assess 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 -
Media used for penalty event detection will be provided by consumer devices. Media may include images and videos captured during the sporting event and shared 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 video transcripts or radio commentary, and logs from related social media feeds captured during the sporting events.
Crowd sentiment will include audio recordings submitted by event attendees in both mono and

stereo formats.

Penalty detection and sentiment -
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.

Advertisements -
During the initial weeks in production, the following was observed:
Ad response rated 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 uncorrelated features.
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 slow.
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 shows 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.
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.
Sampling must guarantee mutual and collective exclusively between 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.
Ad response models must support non-linear boundaries of features.
The ad propensity model uses a cut threshold is 0.45 and retrains occur if weighted Kappa deviated 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:

QuestionYou 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.

Correct Answer: *D*
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.
Incorrect Answers:
B: The Pearson correlation coefficient, sometimes called Pearson's R test, is a statistical value that measures the linear relationship between two variables. By examining the coefficient values, you can infer something about the strength of the relationship between the two variables, and whether they are positively correlated or negatively correlated.
C: Spearman's correlation coefficient is designed for use with non-parametric and non-normally distributed data. Spearman's coefficient is a nonparametric measure of statistical dependence between two variables, and is sometimes denoted by the Greek letter rho. The Spearman's coefficient expresses the degree to which two variables are monotonically related. It is also called Spearman rank correlation, because it can be used with ordinal variables.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/fisher-linear-discriminant-analysis https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/compute-linear-correlation
Question #5*Topic 7*

Introductory Info Case study -

Overview -
You are a data scientist in a company that provides data science for professional sporting events.
Models will use 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.
Assess 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 -
Media used for penalty event detection will be provided by consumer devices. Media may include images and videos captured during the sporting event and shared 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 video transcripts or radio commentary, and logs from related social media feeds captured during the sporting events.
Crowd sentiment will include audio recordings submitted by event attendees in both mono and stereo formats.

Penalty detection and sentiment -
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.

Advertisements -
During the initial weeks in production, the following was observed:
Ad response rated 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 uncorrelated features.
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 slow.
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 shows 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.
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.
Sampling must guarantee mutual and collective exclusively between 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.
Ad response models must support non-linear boundaries of features.
The ad propensity model uses a cut threshold is 0.45 and retrains occur if weighted Kappa deviated 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:

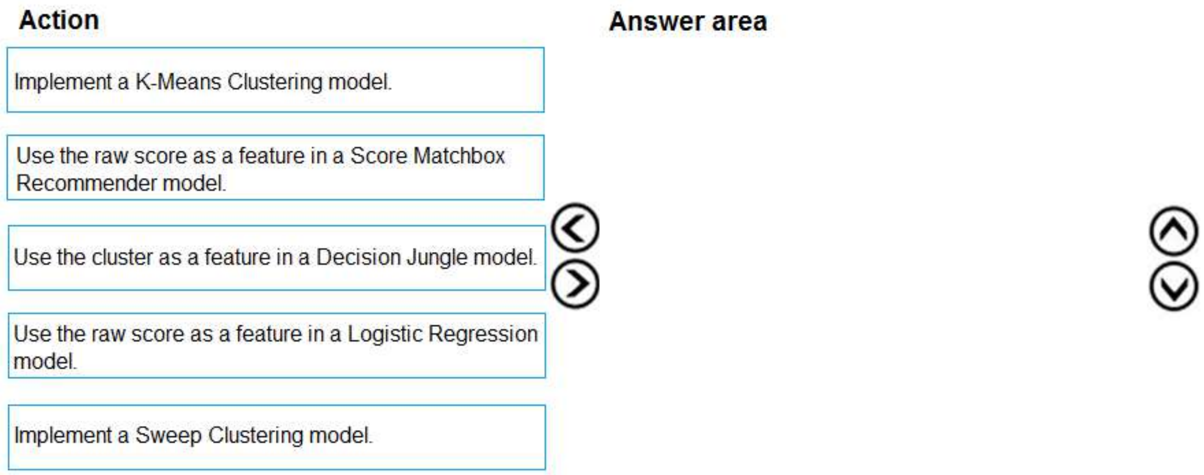
-

		Actual	
		1	0
Predicted	0	1	5
	1	5	1

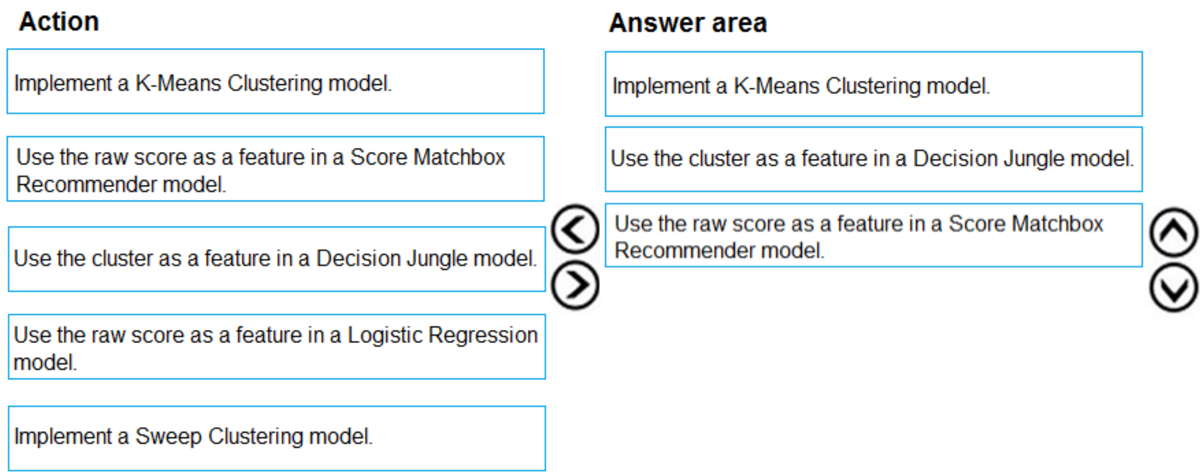
Performance curves of current and proposed cost factor scenarios are shown in the following diagram:

QuestionDRAG DROP -
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.

Select and Place:



Correct Answer:



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.
Reference:
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-recommender
Question #6*Topic 7*

Introductory Info Case study -

Overview -
You are a data scientist in a company that provides data science for professional sporting events. Models will use 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.
Assess 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 -
Media used for penalty event detection will be provided by consumer devices. Media may include images and videos captured during the sporting event and shared 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 video transcripts or radio commentary, and logs from related social media feeds captured during the sporting events.
Crowd sentiment will include audio recordings submitted by event attendees in both mono and stereo formats.

Penalty detection and sentiment -
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.

Advertisements -
During the initial weeks in production, the following was observed:
Ad response rated 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 uncorrelated features.
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 slow.
Audio samples show that the length of a catch phrase varies between 25%-47% depending on region

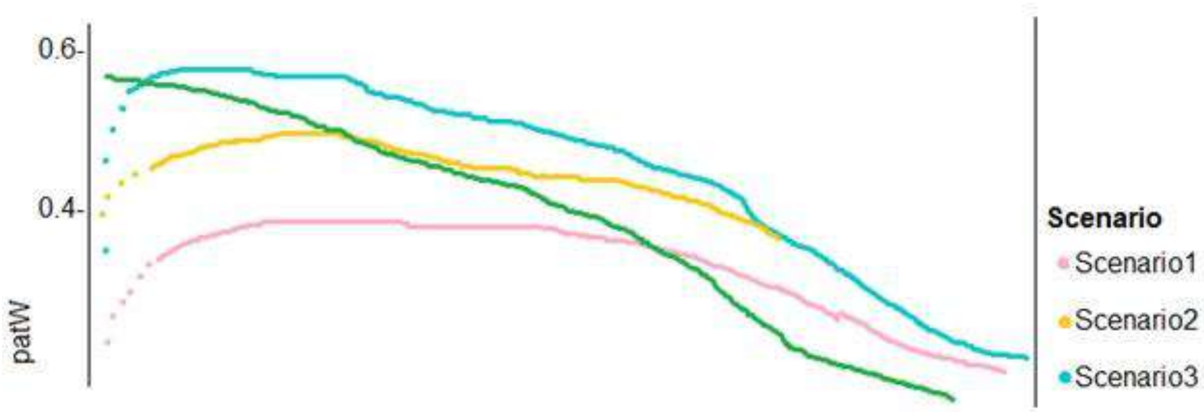
The performance of the global penalty detection models shows lower variance but higher bias when comparing training and validation sets. Before implementing any feature changes, you must confirm the bias and variance using all training and validation cases.
Ad response models must be trained at the beginning of each event and applied during the sporting event.
Market segmentation models must optimize for similar ad response history.
Sampling must guarantee mutual and collective exclusively between 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.
Ad response models must support non-linear boundaries of features.
The ad propensity model uses a cut threshold is 0.45 and retrains occur if weighted Kappa deviated 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:



QuestionDRAG DROP -
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.
Select and Place:

Actions

- Define a cross-entropy function activation.
- Add cost functions for each target state.
- Evaluate the classification error metric.
- Evaluate the distance error metric.
- Add cost functions for each component metric.
- Define a sigmoid loss function activation.

Answer Area

Correct Answer:

Actions

- Define a cross-entropy function activation.
- Add cost functions for each target state.
- Evaluate the classification error metric.
- Evaluate the distance error metric.
- Add cost functions for each component metric.
- Define a sigmoid loss function activation.

Answer Area

- Define a cross-entropy function activation.
- Add cost functions for each target state.
- Evaluate the distance error metric.

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.
Reference:
https://www.analyticsvidhya.com/blog/2018/04/fundamentals-deep-learning-regularization-techniques/
Question #7*Topic 7*

Introductory Info Case study -

Overview -
You are a data scientist in a company that provides data science for professional sporting events.
Models will use 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.
Assess 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 -
Media used for penalty event detection will be provided by consumer devices. Media may include images and videos captured during the sporting event and shared 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 video transcripts or radio commentary, and logs from related social media feeds captured during the sporting events.
Crowd sentiment will include audio recordings submitted by event attendees in both mono and stereo formats.

Penalty detection and sentiment -
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.

Advertisements -
During the initial weeks in production, the following was observed:
Ad response rated 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 uncorrelated features.
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 slow.
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 shows 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.
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.
Sampling must guarantee mutual and collective exclusively between 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.
Ad response models must support non-linear boundaries of features.
The ad propensity model uses a cut threshold is 0.45 and retrains occur if weighted Kappa deviated 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:

QuestionYou 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.

Correct Answer: *A*
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
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/split-data

Question #8*Topic 7*

Introductory Info Case study -

Overview -
You are a data scientist in a company that provides data science for professional sporting events.
Models will use 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.
Assess 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 -
Media used for penalty event detection will be provided by consumer devices. Media may include images and videos captured during the sporting event and shared 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 video transcripts or radio commentary, and logs from related social media feeds captured during the sporting events.
Crowd sentiment will include audio recordings submitted by event attendees in both mono and stereo formats.

Penalty detection and sentiment -
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.

Advertisements -
During the initial weeks in production, the following was observed:
Ad response rated 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 uncorrelated features.
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 slow.
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 shows 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.
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.
Sampling must guarantee mutual and collective exclusively between 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.
Ad response models must support non-linear boundaries of features.
The ad propensity model uses a cut threshold is 0.45 and retrains occur if weighted Kappa deviated 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:

▪

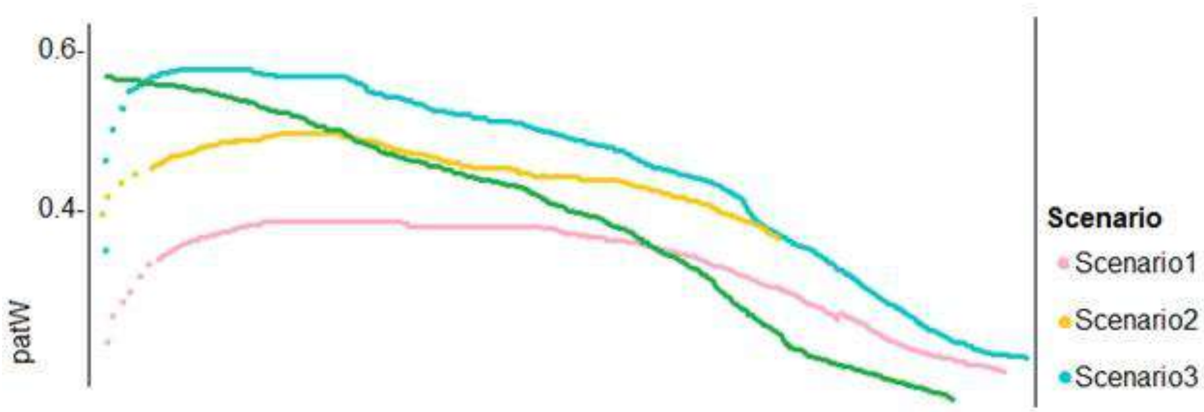
		Actual	
		1	0
Predicted	0	1	5
	1	5	1

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

QuestionYou 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.

Correct Answer: *A*
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%.

Topic 8 - Testlet 2
Question #1*Topic 8*

Introductory Info Case study -
This is a case study. Case studies are not timed separately. You can use as much exam time as you would like to complete each case. However, there may be additional case studies and sections on this exam. You must manage your time to ensure that you are able to complete all questions included on this exam in the time provided.
To answer the questions included in a case study, you will need to reference information that is provided in the case study. Case studies might contain exhibits and other resources that provide more information about the scenario that is described in the case study. Each question is independent of the other questions in this case study.
At the end of this case study, a review screen will appear. This screen allows you to review your answers and to make changes before you move to the next section of the exam. After you begin a new section, you cannot return to this section.

To start the case study -
To display the first question in this case study, click the Next button. Use the buttons in the left pane to explore the content of the case study before you answer the questions. Clicking these buttons displays information such as business requirements, existing environment, and problem statements. If the case study has an
All Information tab, note that the information displayed is identical to the information displayed on the subsequent tabs. When you are ready to answer a question, click the Question button to return to the question.

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. You add both files to Azure Machine Learning Studio as separate datasets to the starting point for an experiment. Both datasets contain 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 a 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

An initial investigation shows 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.

Data issues -

Missing values -
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 datasets also contain 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. 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-parametric statistics to measure relationships.
You must use a feature selection algorithm to analyze the relationship between the MedianValue and AvgRoomsInHouse columns.

Model training -

Permutation Feature Importance -
Given a trained model and a test dataset, you must compute the Permutation Feature Importance scores of feature variables. You must be determined the absolute fit for the model.

Hyperparameters -
You must configure hyperparameters in the model learning process to speed the learning phase. 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, must implement an early stopping criterion on models that provides savings without terminating promising jobs.

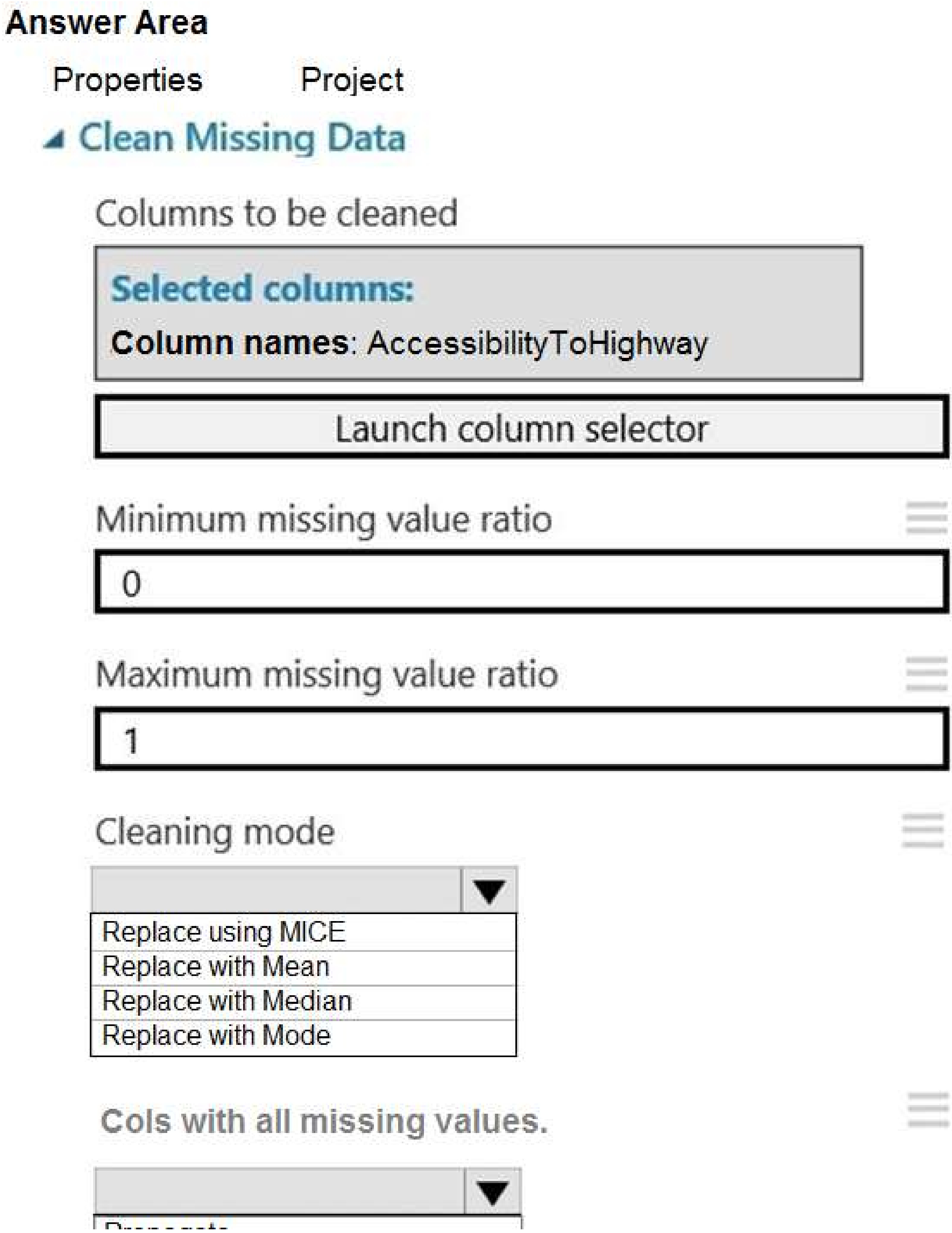
Testing -
You must produce multiple partitions of a dataset based on sampling using the Partition and Sample module in Azure Machine Learning Studio.

Cross-validation -
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. You must complete this task before the data goes through the sampling process.

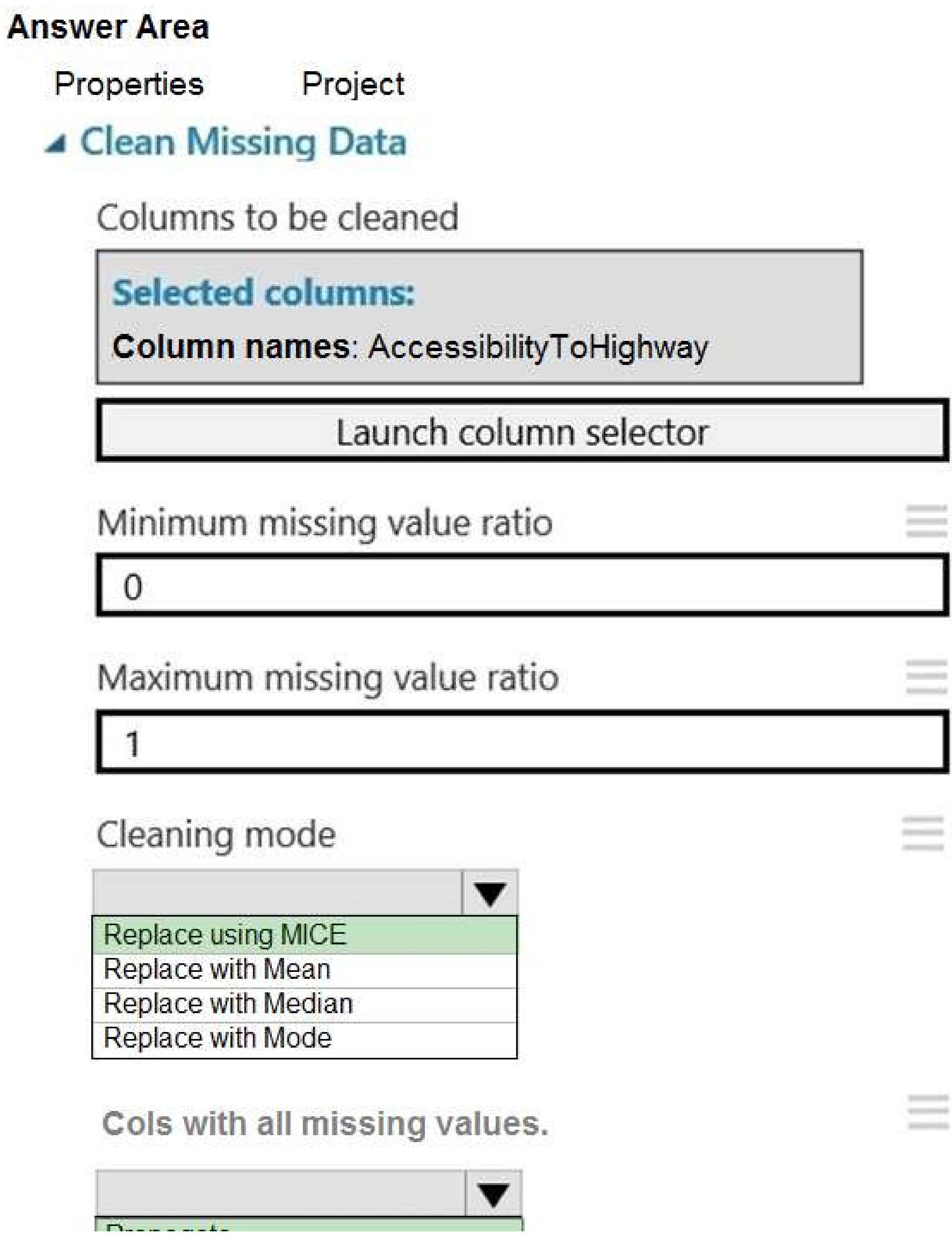
Linear regression module -
When you train a Linear Regression module, you must determine the best features to use in a model. You can choose standard metrics provided to measure performance before and after the feature importance process completes. The distribution of features across multiple training models must be consistent.

Data visualization -
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.**Question**HOTSPOT -
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.

Hot Area:



Correct Answer:



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.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/clean-missing-data
Question #2*Topic 8*

Introductory Info Case study -
This is a case study. Case studies are not timed separately. You can use as much exam time as you would like to complete each case. However, there may be additional case studies and sections on this exam. You must manage your time to ensure that you are able to complete all questions included on this exam in the time provided.
To answer the questions included in a case study, you will need to reference information that is provided in the case study. Case studies might contain exhibits and other resources that provide more information about the scenario that is described in the case study. Each question is independent of the other questions in this case study.
At the end of this case study, a review screen will appear. This screen allows you to review your answers and to make changes before you move to the next section of the exam. After you begin a new section, you cannot return to this section.

To start the case study -
To display the first question in this case study, click the Next button. Use the buttons in the left pane to explore the content of the case study before you answer the questions. Clicking these buttons displays information such as business requirements, existing environment, and problem statements. If the case study has an
All Information tab, note that the information displayed is identical to the information displayed on the subsequent tabs. When you are ready to answer a question, click the Question button to return to the question.

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. You add both files to Azure Machine Learning Studio as separate datasets to the starting point for an experiment. Both datasets contain 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 a 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

An initial investigation shows 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.

Data issues -

Missing values -
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 datasets also contain 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. 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-parametric statistics to measure relationships.
You must use a feature selection algorithm to analyze the relationship between the MedianValue and AvgRoomsInHouse columns.

Model training -

Permutation Feature Importance -
Given a trained model and a test dataset, you must compute the Permutation Feature Importance

scores of feature variables. You must be determined the absolute fit for the model.

Hyperparameters -
You must configure hyperparameters in the model learning process to speed the learning phase. 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, must implement an early stopping criterion on models that provides savings without terminating promising jobs.

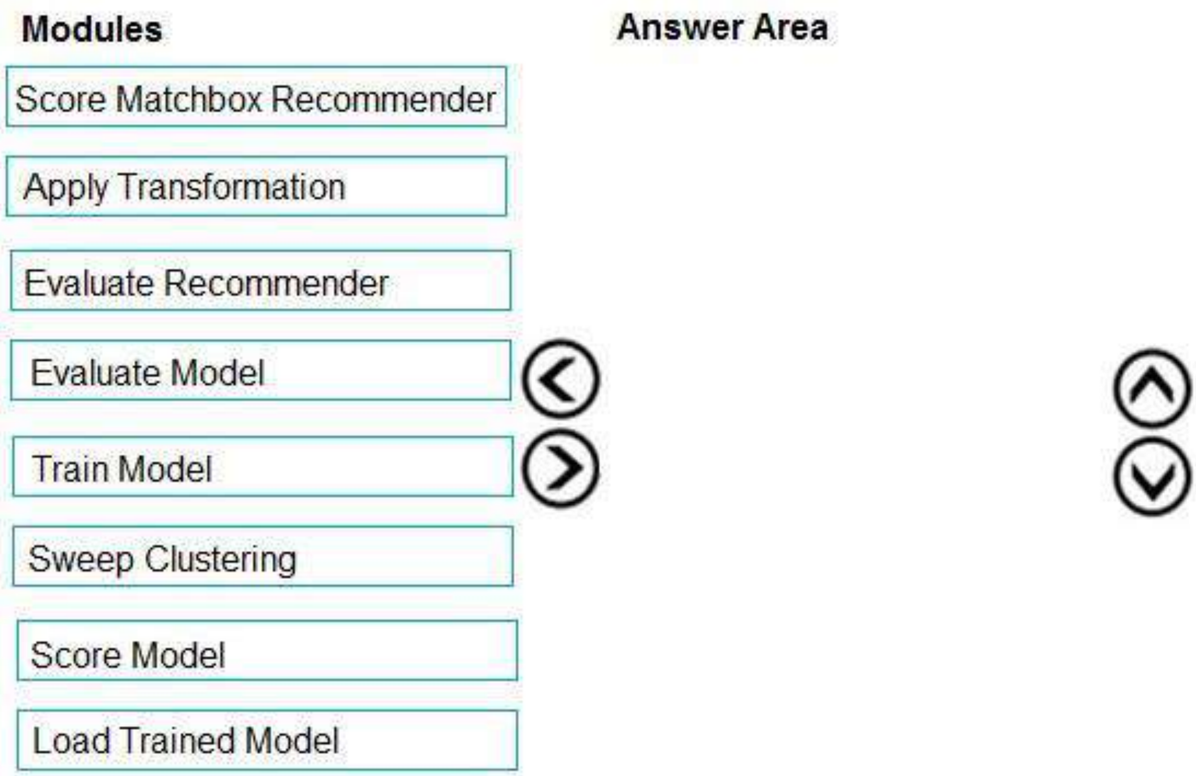
Testing -
You must produce multiple partitions of a dataset based on sampling using the Partition and Sample module in Azure Machine Learning Studio.

Cross-validation -
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. You must complete this task before the data goes through the sampling process.

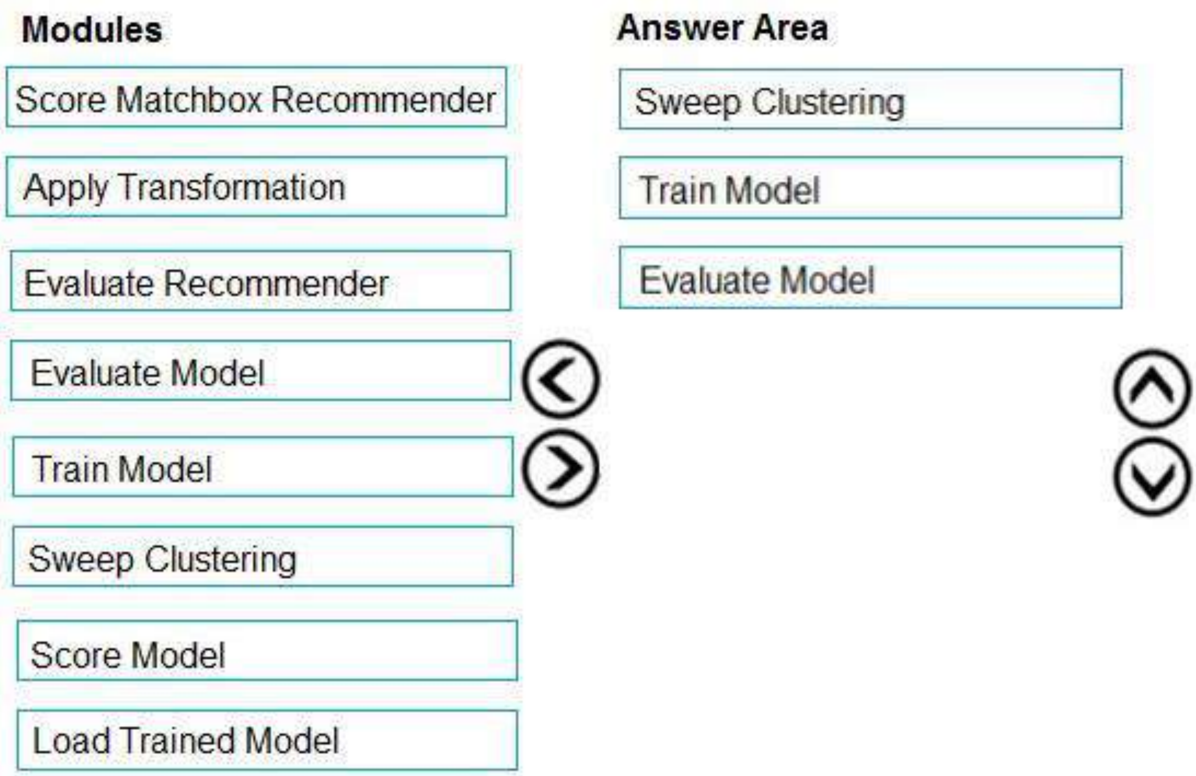
Linear regression module -
When you train a Linear Regression module, you must determine the best features to use in a model. You can choose standard metrics provided to measure performance before and after the feature importance process completes. The distribution of features across multiple training models must be consistent.

Data visualization -
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.**Question**DRAG DROP -
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.

Select and Place:



Correct Answer:



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.
Reference:
http://breaking-bi.blogspot.com/2017/01/azure-machine-learning-model-evaluation.html
Question #3*Topic 8*

Introductory Info Case study -
This is a case study. Case studies are not timed separately. You can use as much exam time as you would like to complete each case. However, there may be additional case studies and sections on this exam. You must manage your time to ensure that you are able to complete all questions included on this exam in the time provided.
To answer the questions included in a case study, you will need to reference information that is provided in the case study. Case studies might contain exhibits and other resources that provide more information about the scenario that is described in the case study. Each question is independent of the other questions in this case study.
At the end of this case study, a review screen will appear. This screen allows you to review your answers and to make changes before you move to the next section of the exam. After you begin a new section, you cannot return to this section.

To start the case study -
To display the first question in this case study, click the Next button. Use the buttons in the left pane to explore the content of the case study before you answer the questions. Clicking these buttons displays information such as business requirements, existing environment, and problem statements. If the case study has an
All Information tab, note that the information displayed is identical to the information displayed on the subsequent tabs. When you are ready to answer a question, click the Question button to return to the question.

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. You add both files to Azure Machine Learning Studio as separate datasets to the starting point for an experiment. Both datasets contain 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 a 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

An initial investigation shows 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.

Data issues -

Missing values -
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 datasets also contain 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. 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-parametric statistics to measure relationships.
You must use a feature selection algorithm to analyze the relationship between the MedianValue and AvgRoomsInHouse columns.

Model training -

Permutation Feature Importance -
Given a trained model and a test dataset, you must compute the Permutation Feature Importance

scores of feature variables. You must be determined the absolute fit for the model.

Hyperparameters -
You must configure hyperparameters in the model learning process to speed the learning phase. 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, must 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.

Cross-validation -
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. You must complete this task before the data goes through the sampling process.

Linear regression module -
When you train a Linear Regression module, you must 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. The distribution of features across multiple training models must be consistent.

Data visualization -
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.**Question**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? Each correct answer presents part of the solution.
NOTE: Each correct selection is worth one point.

- A. Create Scatterplot
- B. Summarize Data
- C. Clip Values
- D. Replace Discrete Values
- E. Build Counting Transform

Correct Answer: *ABC*
B: To have a global view, the summarize data module can be used. Add the module and connect it to the data set that needs to be visualized.
A: One way to quickly identify Outliers visually is to create scatter plots.
C: The easiest way to treat the outliers in Azure ML is to use the Clip Values module. It can identify and optionally replace data values that are above or below a specified threshold.

You can use the Clip Values module in Azure Machine Learning Studio, to identify and optionally replace data values that are above or below a specified threshold. This is useful when you want to remove outliers or replace them with a mean, a constant, or other substitute value.
Reference:
https://blogs.msdn.microsoft.com/azuredev/2017/05/27/data-cleansing-tools-in-azure-machine-learning/ https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/clip-values
Question #4*Topic 8*

Introductory Info Case study -
This is a case study. Case studies are not timed separately. You can use as much exam time as you would like to complete each case. However, there may be additional case studies and sections on this exam. You must manage your time to ensure that you are able to complete all questions included on this exam in the time provided.
To answer the questions included in a case study, you will need to reference information that is provided in the case study. Case studies might contain exhibits and other resources that provide more information about the scenario that is described in the case study. Each question is independent of the other questions in this case study.
At the end of this case study, a review screen will appear. This screen allows you to review your answers and to make changes before you move to the next section of the exam. After you begin a new section, you cannot return to this section.

To start the case study -
To display the first question in this case study, click the Next button. Use the buttons in the left pane to explore the content of the case study before you answer the questions. Clicking these buttons displays information such as business requirements, existing environment, and problem statements. If the case study has an
All Information tab, note that the information displayed is identical to the information displayed on the subsequent tabs. When you are ready to answer a question, click the Question button to return to the question.

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. You add both files to Azure Machine Learning Studio as separate datasets to the starting point for an experiment. Both datasets contain 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 a 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

An initial investigation shows 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.

Data issues -

Missing values -
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 datasets also contain 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. 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-parametric statistics to measure relationships.
You must use a feature selection algorithm to analyze the relationship between the MedianValue and AvgRoomsInHouse columns.

Model training -

Permutation Feature Importance -
Given a trained model and a test dataset, you must compute the Permutation Feature Importance

scores of feature variables. You must be determined the absolute fit for the model.

Hyperparameters -
You must configure hyperparameters in the model learning process to speed the learning phase. 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, must 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.

Cross-validation -
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. You must complete this task before the data goes through the sampling process.

Linear regression module -
When you train a Linear Regression module, you must 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. The distribution of features across multiple training models must be consistent.

Data visualization -
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.**Question**HOTSPOT -
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 options in the answer area.

NOTE: Each correct selection is worth one point.
Hot Area:

Answer Area

Properties Project

◢ Partition and Sample

▼
- Assign to Folds
- Sampling
- Head

Partition or sample mode

☐ Use replacement in the partitioning

☑ Randomized split

Random seed

0

▼
- True
- False
- Partition evenly
- Partition with custom partitions

Specify the partitioner method

Partition evenly

Specify number of folds to split evenly into

3

Correct Answer:

Answer Area

Properties Project

◢ Partition and Sample

Assign to Folds
Sampling
Head

Partition or sample mode

☐ Use replacement in the partitioning

☑ Randomized split

Random seed

0

True
False
Partition evenly
Partition with custom partitions

Specify the partitioner method

Partition evenly

Specify number of folds to split evenly into

3

Scenario: Testing -
You must produce multiple partitions of a dataset based on sampling using the Partition and Sample module in Azure Machine Learning Studio.

Box 1: Assign to folds -
Use Assign to folds option when you want to divide the dataset into subsets of the data. This option is also useful when you want to create a custom number of folds for cross-validation, or to split rows into several groups.
Not Head: Use Head mode to get only the first n rows. This option is useful if you want to test a pipeline on a small number of rows, and don't need the data to be balanced or sampled in any way.
Not Sampling: The Sampling option supports simple random sampling or stratified random sampling. This is useful if you want to create a smaller representative sample dataset for testing.

Box 2: Partition evenly -
Specify the partitioner method: Indicate how you want data to be apportioned to each partition, using these options:
⇨ Partition evenly: Use this option to place an equal number of rows in each partition. To specify the number of output partitions, type a whole number in the
Specify number of folds to split evenly into text box.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/algorithm-module-reference/partition-and-sample
Question #5*Topic 8*

Introductory Info Case study -
This is a case study. Case studies are not timed separately. You can use as much exam time as you would like to complete each case. However, there may be additional case studies and sections on this exam. You must manage your time to ensure that you are able to complete all questions included on this exam in the time provided.
To answer the questions included in a case study, you will need to reference information that is provided in the case study. Case studies might contain exhibits and other resources that provide more information about the scenario that is described in the case study. Each question is independent of the other questions in this case study.
At the end of this case study, a review screen will appear. This screen allows you to review your answers and to make changes before you move to the next section of the exam. After you begin a new section, you cannot return to this section.

To start the case study -
To display the first question in this case study, click the Next button. Use the buttons in the left pane to explore the content of the case study before you answer the questions. Clicking these buttons displays information such as business requirements, existing environment, and problem statements. If the case study has an
All Information tab, note that the information displayed is identical to the information displayed on the subsequent tabs. When you are ready to answer a question, click the Question button to return to the question.

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. You add both files to Azure Machine Learning Studio as separate datasets to the starting point for an experiment. Both datasets contain 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 a 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

An initial investigation shows 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.

Data issues -

Missing values -
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 datasets also contain 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. 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-parametric statistics to measure relationships.
You must use a feature selection algorithm to analyze the relationship between the MedianValue

and AvgRoomsInHouse columns.

Model training -

Permutation Feature Importance -
Given a trained model and a test dataset, you must compute the Permutation Feature Importance scores of feature variables. You must be determined the absolute fit for the model.

Hyperparameters -
You must configure hyperparameters in the model learning process to speed the learning phase. 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, must implement an early stopping criterion on models that provides savings without terminating promising jobs.

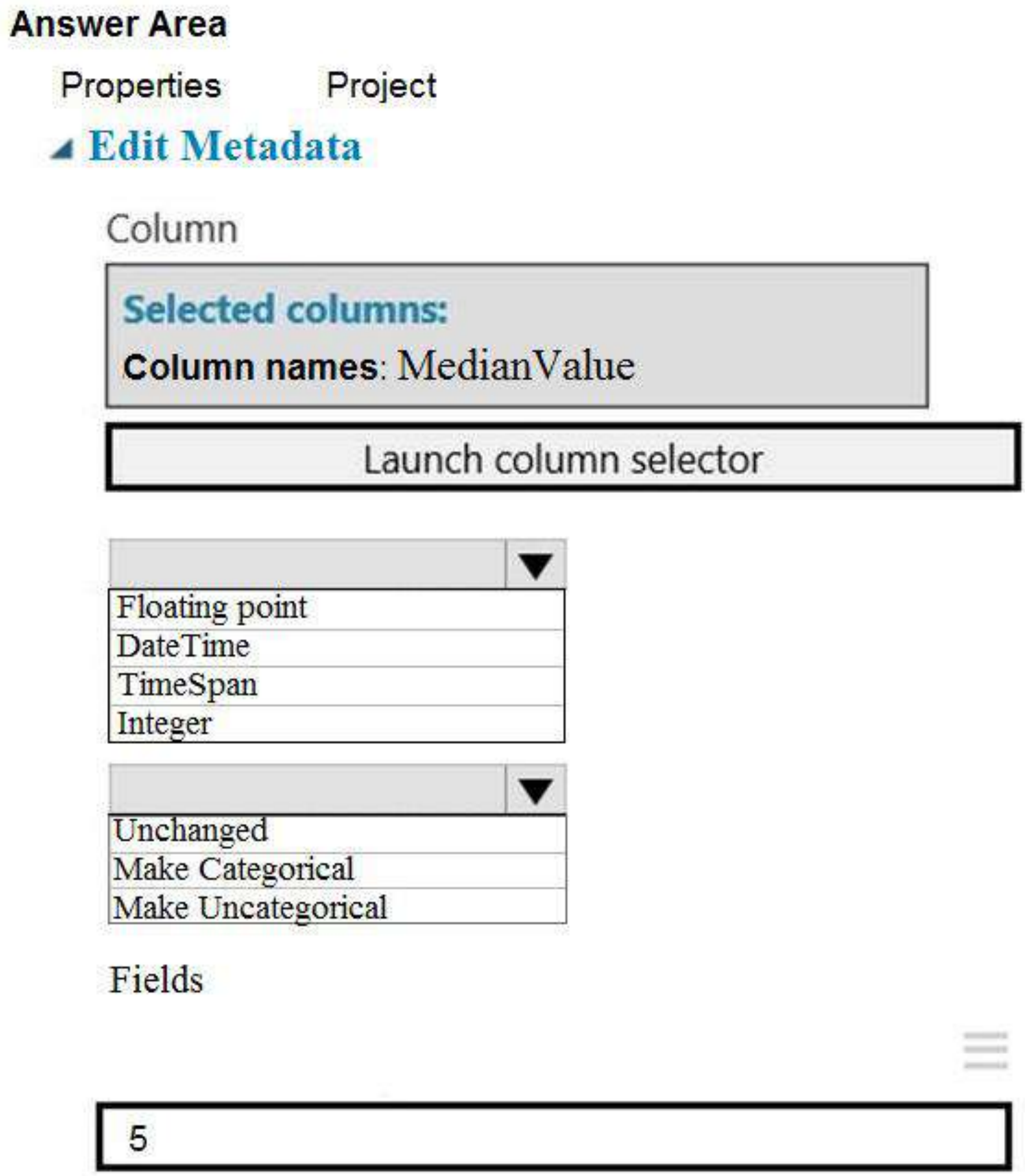
Testing -
You must produce multiple partitions of a dataset based on sampling using the Partition and Sample module in Azure Machine Learning Studio.

Cross-validation -
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. You must complete this task before the data goes through the sampling process.

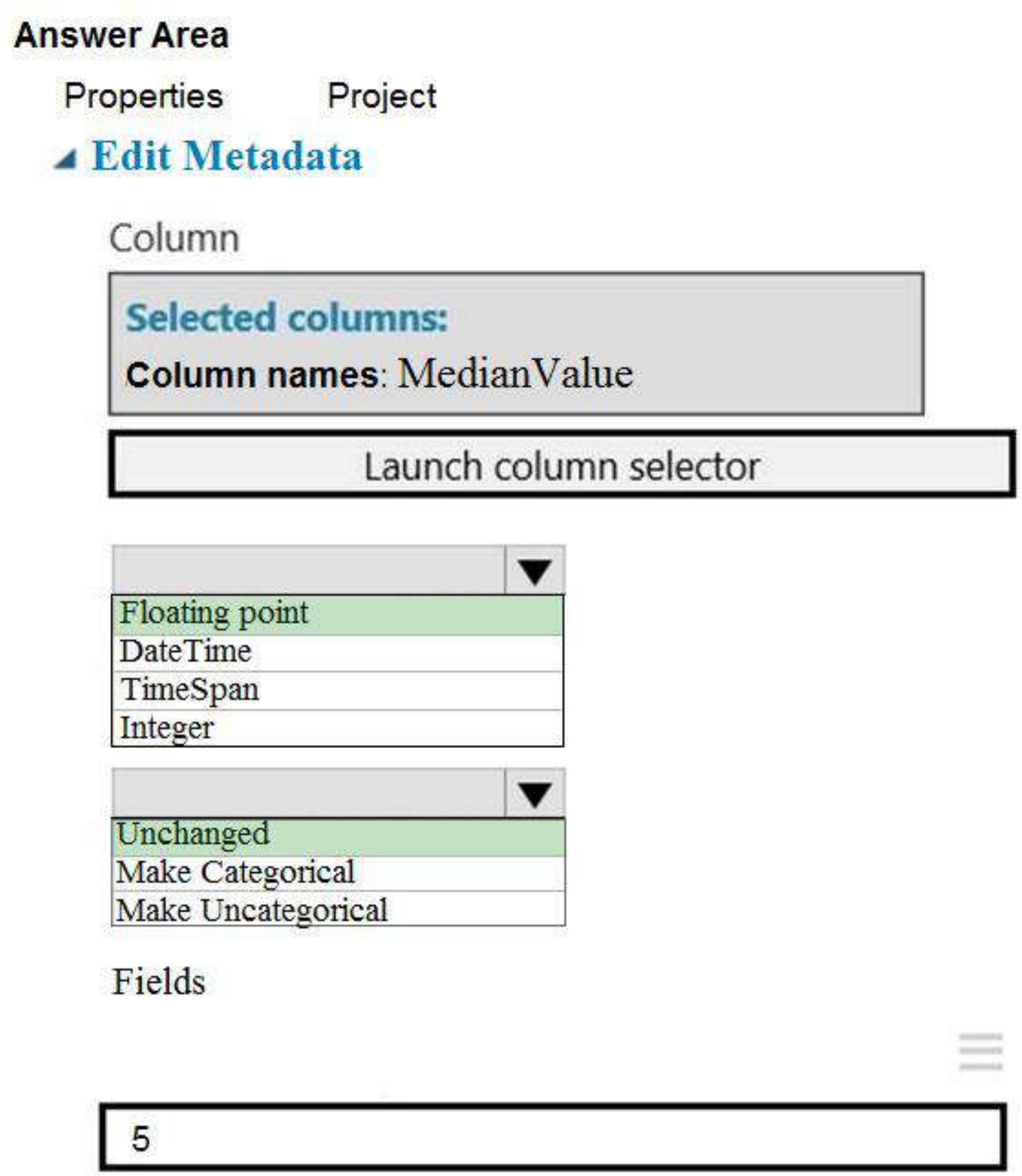
Linear regression module -
When you train a Linear Regression module, you must determine the best features to use in a model. You can choose standard metrics provided to measure performance before and after the feature importance process completes. The distribution of features across multiple training models must be consistent.

Data visualization -
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.**Question**HOTSPOT -
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.

Hot Area:



Correct Answer:



Box 1: Floating point -
Need floating point for Median values.
Scenario: An initial investigation shows 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.

Box 2: Unchanged -
Note: Select the Categorical option to specify that the values in the selected columns should be treated as categories.
For example, you might have a column that contains the numbers 0,1 and 2, but know that the numbers actually mean "Smoker", "Non smoker" and "Unknown". In that case, by flagging the

column as categorical you can ensure that the values are not used in numeric calculations, only to group data.
Question #6*Topic 8*

Introductory Info Case study -
This is a case study. Case studies are not timed separately. You can use as much exam time as you would like to complete each case. However, there may be additional case studies and sections on this exam. You must manage your time to ensure that you are able to complete all questions included on this exam in the time provided.
To answer the questions included in a case study, you will need to reference information that is provided in the case study. Case studies might contain exhibits and other resources that provide more information about the scenario that is described in the case study. Each question is independent of the other questions in this case study.
At the end of this case study, a review screen will appear. This screen allows you to review your answers and to make changes before you move to the next section of the exam. After you begin a new section, you cannot return to this section.

To start the case study -
To display the first question in this case study, click the Next button. Use the buttons in the left pane to explore the content of the case study before you answer the questions. Clicking these buttons displays information such as business requirements, existing environment, and problem statements. If the case study has an
All Information tab, note that the information displayed is identical to the information displayed on the subsequent tabs. When you are ready to answer a question, click the Question button to return to the question.

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. You add both files to Azure Machine Learning Studio as separate datasets to the starting point for an experiment. Both datasets contain 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 a 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

An initial investigation shows 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.

Data issues -

Missing values -
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 datasets also contain 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. 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-parametric statistics to measure relationships.
You must use a feature selection algorithm to analyze the relationship between the MedianValue and AvgRoomsInHouse columns.

Model training -

Permutation Feature Importance -
Given a trained model and a test dataset, you must compute the Permutation Feature Importance

scores of feature variables. You must be determined the absolute fit for the model.

Hyperparameters -
You must configure hyperparameters in the model learning process to speed the learning phase. 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, must implement an early stopping criterion on models that provides savings without terminating promising jobs.

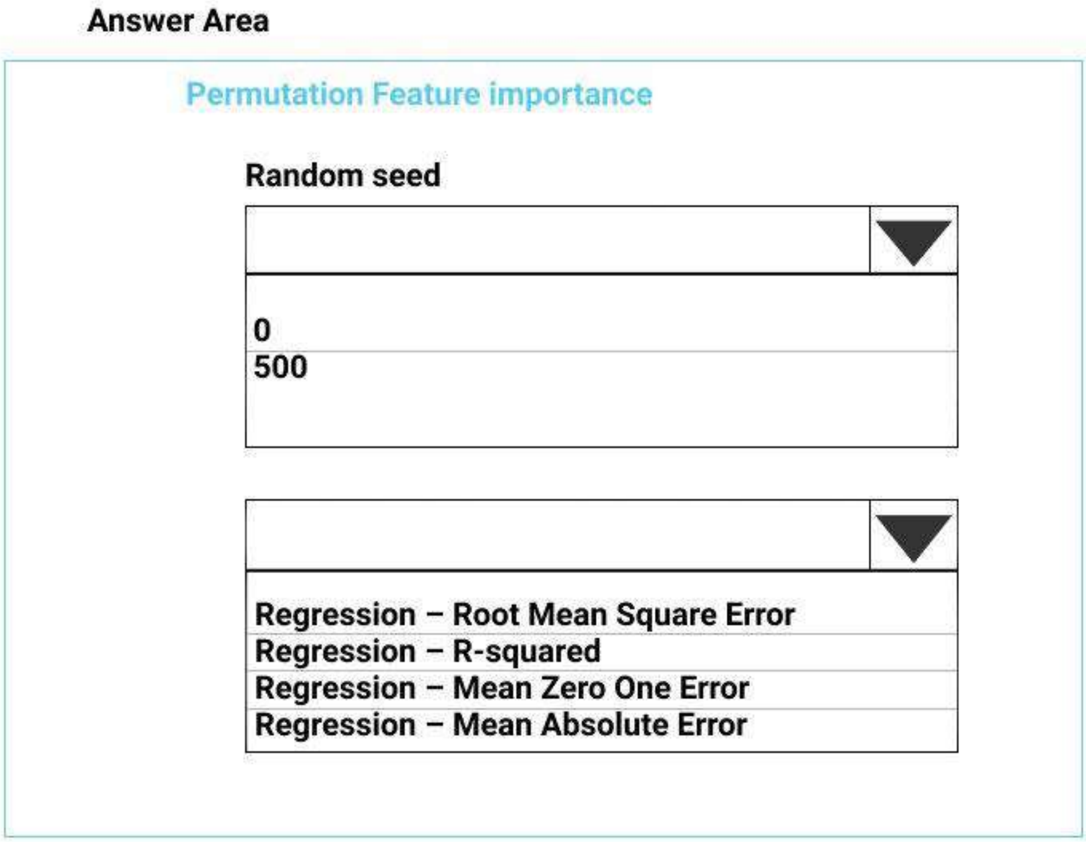
Testing -
You must produce multiple partitions of a dataset based on sampling using the Partition and Sample module in Azure Machine Learning Studio.

Cross-validation -
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. You must complete this task before the data goes through the sampling process.

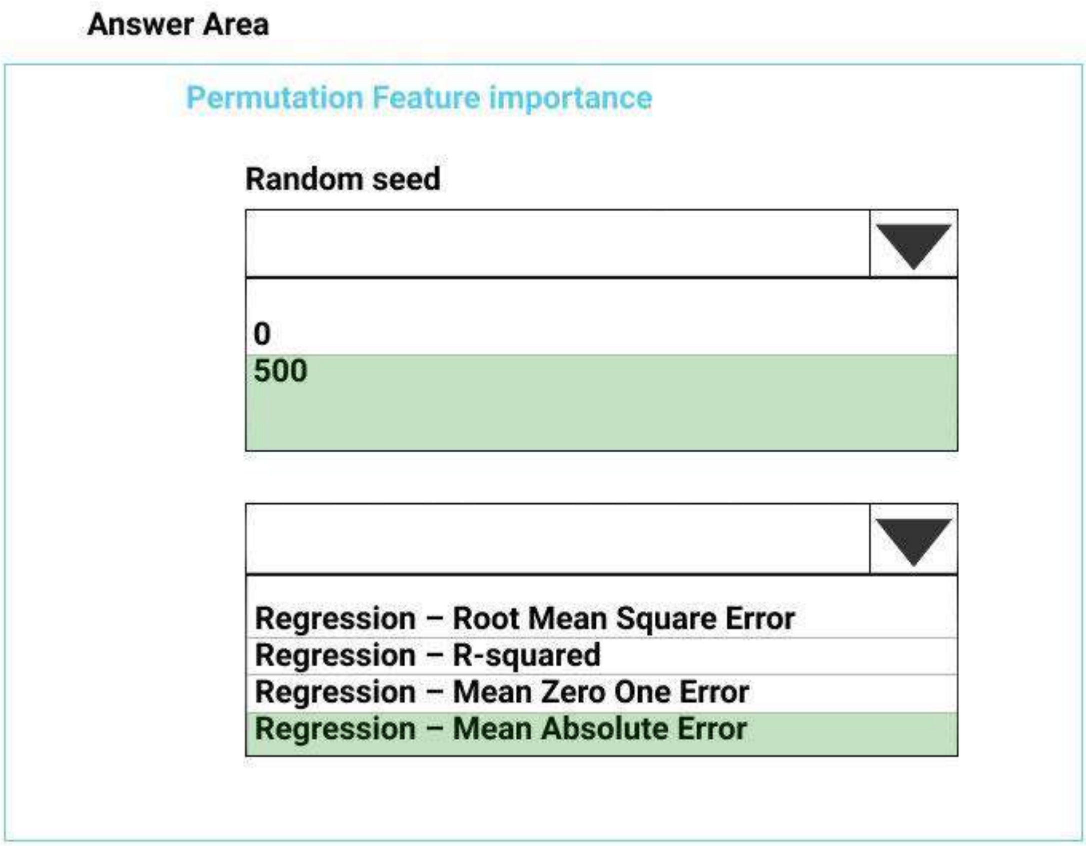
Linear regression module -
When you train a Linear Regression module, you must determine the best features to use in a model. You can choose standard metrics provided to measure performance before and after the feature importance process completes. The distribution of features across multiple training models must be consistent.

Data visualization -
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.**Question**HOTSPOT -
You need to configure the Permutation Feature Importance module for the model training 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.

Hot Area:



Correct Answer:



Box 1: 500 -
For Random seed, type a value to use as seed for randomization. If you specify 0 (the default), a number is generated based on the system clock.
A seed value is optional, but you should provide a value if you want reproducibility across runs of the same experiment.
Here we must replicate the findings.

Box 2: Mean Absolute Error -
Scenario: Given a trained model and a test dataset, you must 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.
Regression. Choose one of the following: Precision, Recall, Mean Absolute Error, Root Mean Squared Error, Relative Absolute Error, Relative Squared Error,

Coefficient of Determination -
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/permutation-feature-importance
Question #7*Topic 8*

Introductory Info Case study -
This is a case study. Case studies are not timed separately. You can use as much exam time as you would like to complete each case. However, there may be additional case studies and sections on this exam. You must manage your time to ensure that you are able to complete all questions

included on this exam in the time provided.
To answer the questions included in a case study, you will need to reference information that is provided in the case study. Case studies might contain exhibits and other resources that provide more information about the scenario that is described in the case study. Each question is independent of the other questions in this case study.
At the end of this case study, a review screen will appear. This screen allows you to review your answers and to make changes before you move to the next section of the exam. After you begin a new section, you cannot return to this section.

To start the case study -
To display the first question in this case study, click the Next button. Use the buttons in the left pane to explore the content of the case study before you answer the questions. Clicking these buttons displays information such as business requirements, existing environment, and problem statements. If the case study has an
All Information tab, note that the information displayed is identical to the information displayed on the subsequent tabs. When you are ready to answer a question, click the Question button to return to the question.

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. You add both files to Azure Machine Learning Studio as separate datasets to the starting point for an experiment. Both datasets contain 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 a 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

An initial investigation shows 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.

Data issues -

Missing values -
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 datasets also contain 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. 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-parametric statistics to measure relationships.
You must use a feature selection algorithm to analyze the relationship between the MedianValue and AvgRoomsInHouse columns.

Model training -

Permutation Feature Importance -
Given a trained model and a test dataset, you must compute the Permutation Feature Importance scores of feature variables. You must be determined the absolute fit for the model.

Hyperparameters -
You must configure hyperparameters in the model learning process to speed the learning phase. 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, must 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.

Cross-validation -
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. You must complete this task before the data goes through the sampling process.

Linear regression module -
When you train a Linear Regression module, you must 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. The distribution of features across multiple training models must be consistent.

Data visualization -
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.**Question**You need to select a feature extraction method.
Which method should you use?

- A. Mutual information
- B. Pearson's correlation
- C. Spearman correlation
- D. Fisher Linear Discriminant Analysis

Correct Answer: *C*
Spearman's rank correlation coefficient assesses how well the relationship between two variables can be described using a monotonic function.
Note: Both Spearman's and Kendall's can be formulated as special cases of a more general correlation coefficient, and they are both appropriate in this scenario.
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.
Incorrect Answers:
B: The Spearman correlation between two variables is equal to the Pearson correlation between the rank values of those two variables; while Pearson's correlation assesses linear relationships, Spearman's correlation assesses monotonic relationships (whether linear or not).
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/feature-selection-modules
Question #8*Topic 8*

Introductory Info Case study -
This is a case study. Case studies are not timed separately. You can use as much exam time as you would like to complete each case. However, there may be additional case studies and sections on this exam. You must manage your time to ensure that you are able to complete all questions included on this exam in the time provided.
To answer the questions included in a case study, you will need to reference information that is provided in the case study. Case studies might contain exhibits and other resources that provide more information about the scenario that is described in the case study. Each question is independent of the other questions in this case study.
At the end of this case study, a review screen will appear. This screen allows you to review your answers and to make changes before you move to the next section of the exam. After you begin a new section, you cannot return to this section.

To start the case study -
To display the first question in this case study, click the Next button. Use the buttons in the left pane

to explore the content of the case study before you answer the questions. Clicking these buttons displays information such as business requirements, existing environment, and problem statements. If the case study has an
All Information tab, note that the information displayed is identical to the information displayed on the subsequent tabs. When you are ready to answer a question, click the Question button to return to the question.

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. You add both files to Azure Machine Learning Studio as separate datasets to the starting point for an experiment. Both datasets contain 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 a 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

An initial investigation shows 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.

Data issues -

Missing values -
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 datasets also contain 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. 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-parametric statistics to measure relationships.
You must use a feature selection algorithm to analyze the relationship between the MedianValue and AvgRoomsInHouse columns.

Model training -

Permutation Feature Importance -
Given a trained model and a test dataset, you must compute the Permutation Feature Importance scores of feature variables. You must be determined the absolute fit for the model.

Hyperparameters -
You must configure hyperparameters in the model learning process to speed the learning phase. 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, must implement an early stopping criterion on models that provides savings without terminating promising jobs.

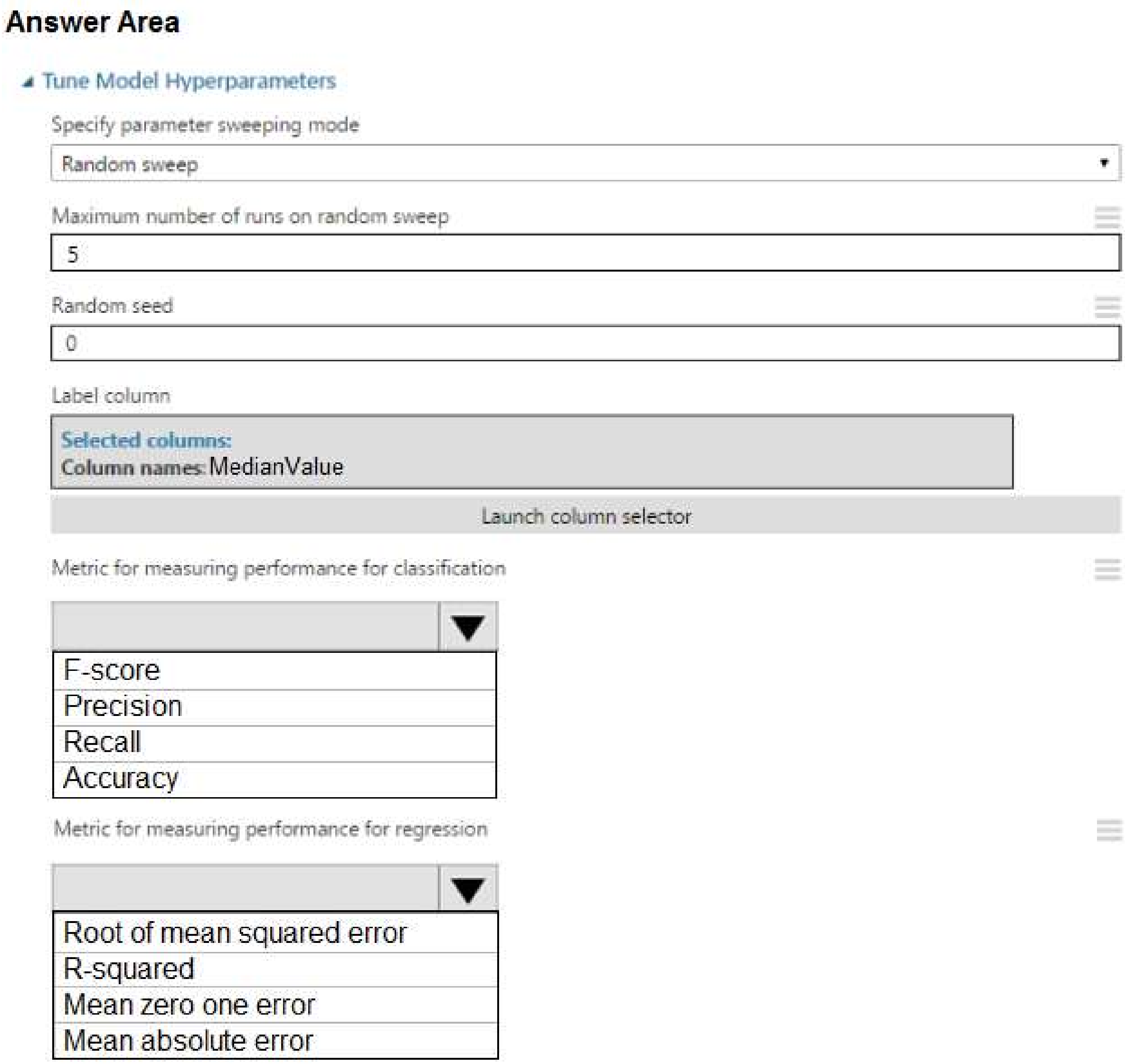
Testing -
You must produce multiple partitions of a dataset based on sampling using the Partition and Sample module in Azure Machine Learning Studio.

Cross-validation -
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. You must complete this task before the data goes through the sampling process.

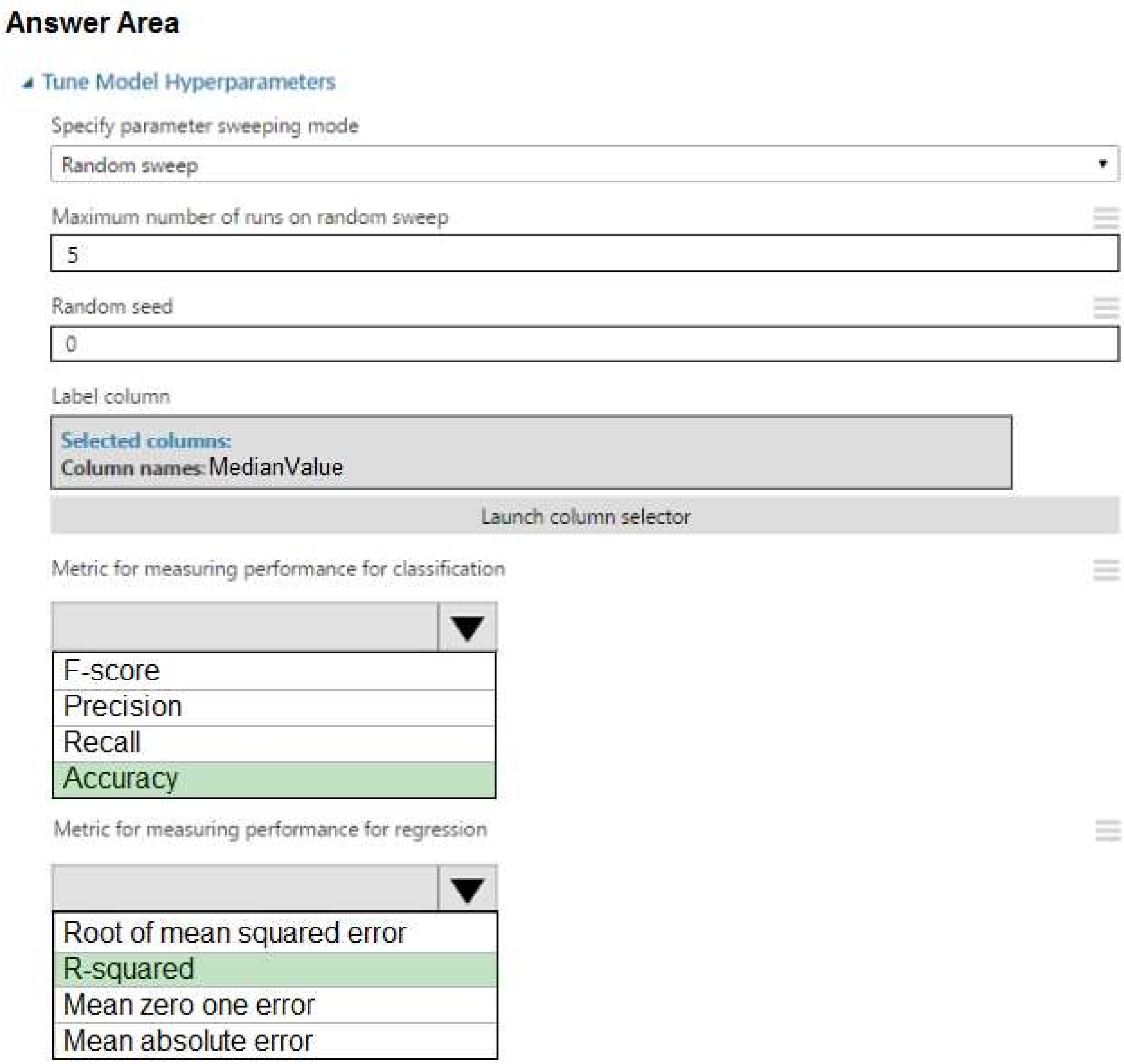
Linear regression module -
When you train a Linear Regression module, you must determine the best features to use in a model. You can choose standard metrics provided to measure performance before and after the feature importance process completes. The distribution of features across multiple training models must be consistent.

Data visualization -
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.**Question**HOTSPOT -
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.
Hot Area:



Correct Answer:



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
Question #9*Topic 8*

Introductory Info Case study -
This is a case study. Case studies are not timed separately. You can use as much exam time as you would like to complete each case. However, there may be additional case studies and sections on this exam. You must manage your time to ensure that you are able to complete all questions included on this exam in the time provided.
To answer the questions included in a case study, you will need to reference information that is provided in the case study. Case studies might contain exhibits and other resources that provide more information about the scenario that is described in the case study. Each question is independent of the other questions in this case study.
At the end of this case study, a review screen will appear. This screen allows you to review your answers and to make changes before you move to the next section of the exam. After you begin a

new section, you cannot return to this section.

To start the case study -
To display the first question in this case study, click the Next button. Use the buttons in the left pane to explore the content of the case study before you answer the questions. Clicking these buttons displays information such as business requirements, existing environment, and problem statements. If the case study has an
All Information tab, note that the information displayed is identical to the information displayed on the subsequent tabs. When you are ready to answer a question, click the Question button to return to the question.

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. You add both files to Azure Machine Learning Studio as separate datasets to the starting point for an experiment. Both datasets contain 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 a 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

An initial investigation shows 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.

Data issues -

Missing values -
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 datasets also contain 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. 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-parametric statistics to measure relationships.
You must use a feature selection algorithm to analyze the relationship between the MedianValue and AvgRoomsInHouse columns.

Model training -

Permutation Feature Importance -
Given a trained model and a test dataset, you must compute the Permutation Feature Importance scores of feature variables. You must be determined the absolute fit for the model.

Hyperparameters -
You must configure hyperparameters in the model learning process to speed the learning phase. 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, must implement an early stopping criterion on models that provides savings without terminating promising jobs.

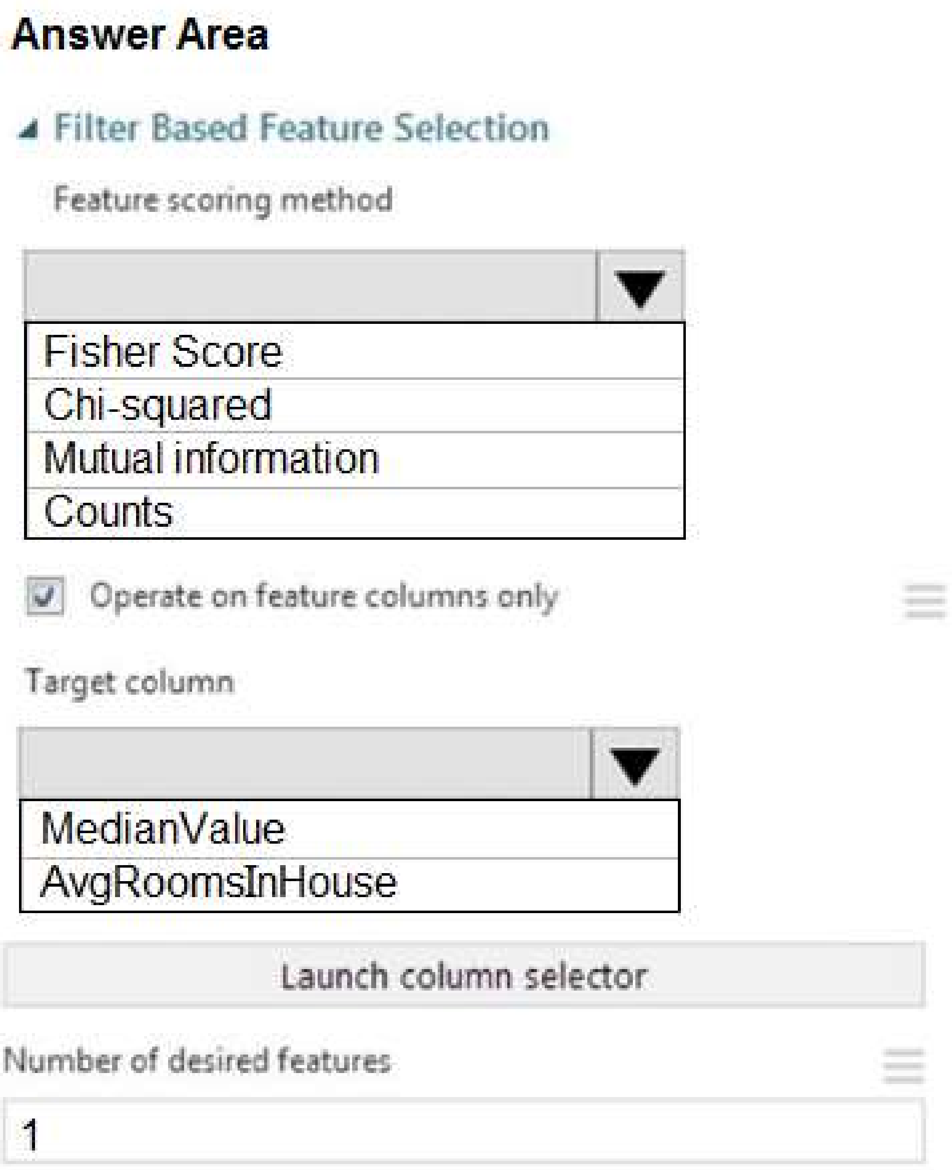
Testing -
You must produce multiple partitions of a dataset based on sampling using the Partition and Sample module in Azure Machine Learning Studio.

Cross-validation -
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. You must complete this task before the data goes through the sampling process.

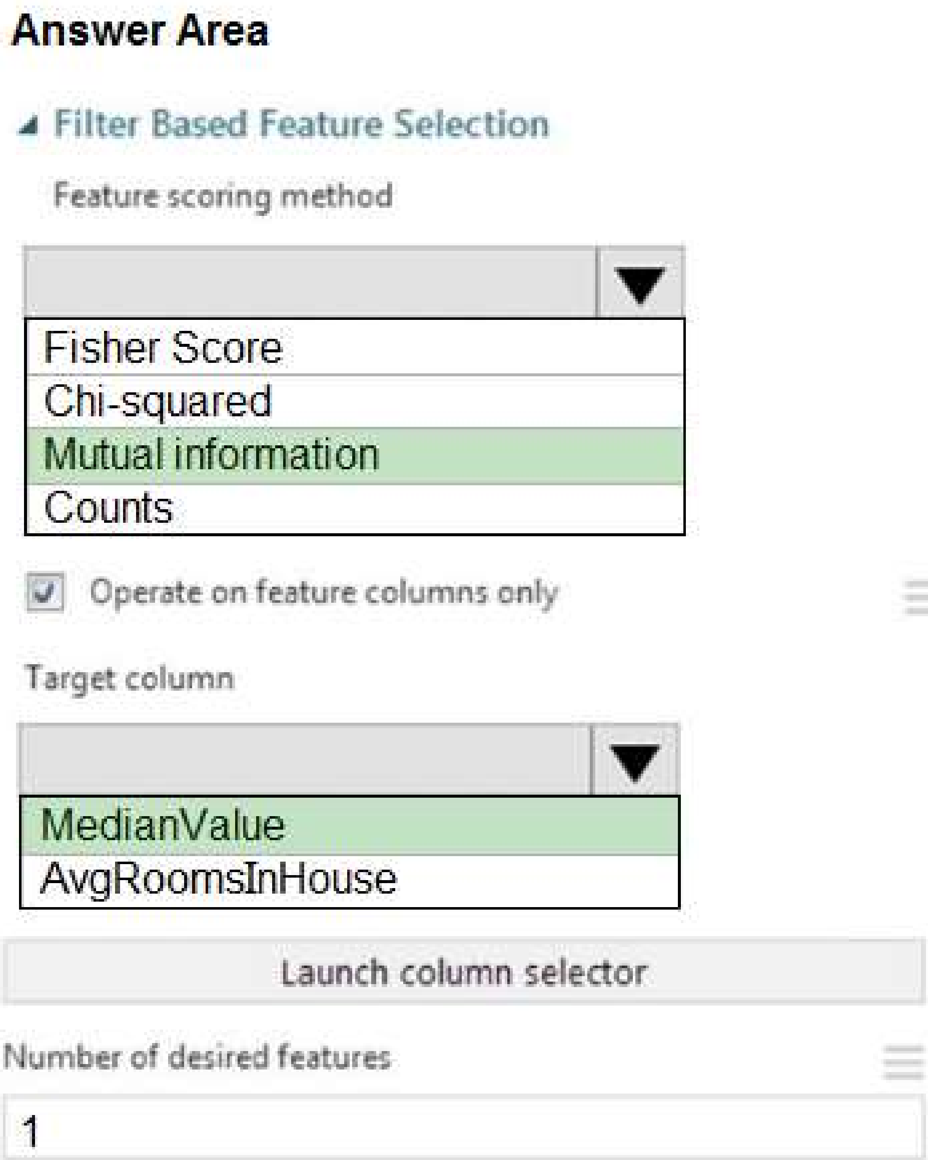
Linear regression module -
When you train a Linear Regression module, you must determine the best features to use in a model. You can choose standard metrics provided to measure performance before and after the feature importance process completes. The distribution of features across multiple training models must be consistent.

Data visualization -

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.**Question**HOTSPOT -
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.
Hot Area:



Correct



Answer:
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.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/filter-based-feature-selection
Question #10*Topic 8*

Introductory Info Case study -
This is a case study. Case studies are not timed separately. You can use as much exam time as you

would like to complete each case. However, there may be additional case studies and sections on this exam. You must manage your time to ensure that you are able to complete all questions included on this exam in the time provided.
To answer the questions included in a case study, you will need to reference information that is provided in the case study. Case studies might contain exhibits and other resources that provide more information about the scenario that is described in the case study. Each question is independent of the other questions in this case study.
At the end of this case study, a review screen will appear. This screen allows you to review your answers and to make changes before you move to the next section of the exam. After you begin a new section, you cannot return to this section.

To start the case study -
To display the first question in this case study, click the Next button. Use the buttons in the left pane to explore the content of the case study before you answer the questions. Clicking these buttons displays information such as business requirements, existing environment, and problem statements. If the case study has an
All Information tab, note that the information displayed is identical to the information displayed on the subsequent tabs. When you are ready to answer a question, click the Question button to return to the question.

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. You add both files to Azure Machine Learning Studio as separate datasets to the starting point for an experiment. Both datasets contain 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 a 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

An initial investigation shows 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.

Data issues -

Missing values -
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 datasets also contain 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. 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-parametric statistics to measure relationships.
You must use a feature selection algorithm to analyze the relationship between the MedianValue and AvgRoomsInHouse columns.

Model training -

Permutation Feature Importance -
Given a trained model and a test dataset, you must compute the Permutation Feature Importance scores of feature variables. You must be determined the absolute fit for the model.

Hyperparameters -
You must configure hyperparameters in the model learning process to speed the learning phase. 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, must 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.

Cross-validation -
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. You must complete this task before the data goes through the

sampling process.

Linear regression module -
When you train a Linear Regression module, you must 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. The distribution of features across multiple training models must be consistent.

Data visualization -
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.**Question**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

Correct Answer: *C*
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.
Note: Both Spearman's and Kendall's can be formulated as special cases of a more general correlation coefficient, and they are both appropriate in this scenario.
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.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/feature-selection-modules
Question #11*Topic 8*

Introductory Info Case study -
This is a case study. Case studies are not timed separately. You can use as much exam time as you would like to complete each case. However, there may be additional case studies and sections on this exam. You must manage your time to ensure that you are able to complete all questions included on this exam in the time provided.
To answer the questions included in a case study, you will need to reference information that is provided in the case study. Case studies might contain exhibits and other resources that provide more information about the scenario that is described in the case study. Each question is independent of the other questions in this case study.
At the end of this case study, a review screen will appear. This screen allows you to review your answers and to make changes before you move to the next section of the exam. After you begin a new section, you cannot return to this section.

To start the case study -
To display the first question in this case study, click the Next button. Use the buttons in the left pane

to explore the content of the case study before you answer the questions. Clicking these buttons displays information such as business requirements, existing environment, and problem statements. If the case study has an
All Information tab, note that the information displayed is identical to the information displayed on the subsequent tabs. When you are ready to answer a question, click the Question button to return to the question.

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. You add both files to Azure Machine Learning Studio as separate datasets to the starting point for an experiment. Both datasets contain 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 a 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

An initial investigation shows 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.

Data issues -

Missing values -
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 datasets also contain 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. 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-parametric statistics to measure relationships.
You must use a feature selection algorithm to analyze the relationship between the MedianValue and AvgRoomsInHouse columns.

Model training -

Permutation Feature Importance -
Given a trained model and a test dataset, you must compute the Permutation Feature Importance scores of feature variables. You must be determined the absolute fit for the model.

Hyperparameters -
You must configure hyperparameters in the model learning process to speed the learning phase. 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, must 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.

Cross-validation -
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. You must complete this task before the data goes through the sampling process.

Linear regression module -
When you train a Linear Regression module, you must 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. The distribution of features across multiple training models must be consistent.

Data visualization -
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.**Question**DRAG DROP -

You need to implement an early stopping criteria policy for model training.

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.

Select and Place:

Code segments

```
early_termination_policy =
TruncationSelectionPolicy(evaluation_interval=1,
truncation_percentage=20, delay_evaluation=5)
```

```
import TruncationSelectionPolicy
```

```
from azureml.train.hyperdrive
```

```
import BanditPolicy
```

```
early_termination_policy = BanditPolicy
(slack_factor = 0.1, evaluation_interval=1,
delay_evaluation=5)
```

Answer Area

Correct Answer:

Code segments

```
early_termination_policy =
TruncationSelectionPolicy(evaluation_interval=1,
truncation_percentage=20, delay_evaluation=5)
```

```
import TruncationSelectionPolicy
```

```
from azureml.train.hyperdrive
```

```
import BanditPolicy
```

```
early_termination_policy = BanditPolicy
(slack_factor = 0.1, evaluation_interval=1,
delay_evaluation=5)
```

Answer Area

```
from azureml.train.hyperdrive
```

```
import TruncationSelectionPolicy
```

```
early_termination_policy =
TruncationSelectionPolicy(evaluation_interval=1,
truncation_percentage=20, delay_evaluation=5)
```

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)

Incorrect Answers:

Bandit is a termination policy based on slack factor/slack amount and evaluation interval. The policy

early terminates any runs where the primary metric is not within the specified slack factor / slack amount with respect to the best performing training run.
Example:
from azureml.train.hyperdrive import BanditPolicy
early_termination_policy = BanditPolicy(slack_factor = 0.1, evaluation_interval=1, delay_evaluation=5
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/service/how-to-tune-hyperparameters

Question #12*Topic 8*

Introductory Info Case study -
This is a case study. Case studies are not timed separately. You can use as much exam time as you would like to complete each case. However, there may be additional case studies and sections on this exam. You must manage your time to ensure that you are able to complete all questions included on this exam in the time provided.
To answer the questions included in a case study, you will need to reference information that is provided in the case study. Case studies might contain exhibits and other resources that provide more information about the scenario that is described in the case study. Each question is independent of the other questions in this case study.
At the end of this case study, a review screen will appear. This screen allows you to review your answers and to make changes before you move to the next section of the exam. After you begin a new section, you cannot return to this section.

To start the case study -
To display the first question in this case study, click the Next button. Use the buttons in the left pane to explore the content of the case study before you answer the questions. Clicking these buttons displays information such as business requirements, existing environment, and problem statements. If the case study has an
All Information tab, note that the information displayed is identical to the information displayed on the subsequent tabs. When you are ready to answer a question, click the Question button to return to the question.

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. You add both files to Azure Machine Learning Studio as separate datasets to the starting point for an experiment. Both datasets contain 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 a 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

An initial investigation shows 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.

Data issues -

Missing values -
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 datasets also contain 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. 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-parametric statistics to measure relationships.
You must use a feature selection algorithm to analyze the relationship between the MedianValue and AvgRoomsInHouse columns.

Model training -

Permutation Feature Importance -
Given a trained model and a test dataset, you must compute the Permutation Feature Importance

scores of feature variables. You must be determined the absolute fit for the model.

Hyperparameters -
You must configure hyperparameters in the model learning process to speed the learning phase. 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, must implement an early stopping criterion on models that provides savings without terminating promising jobs.

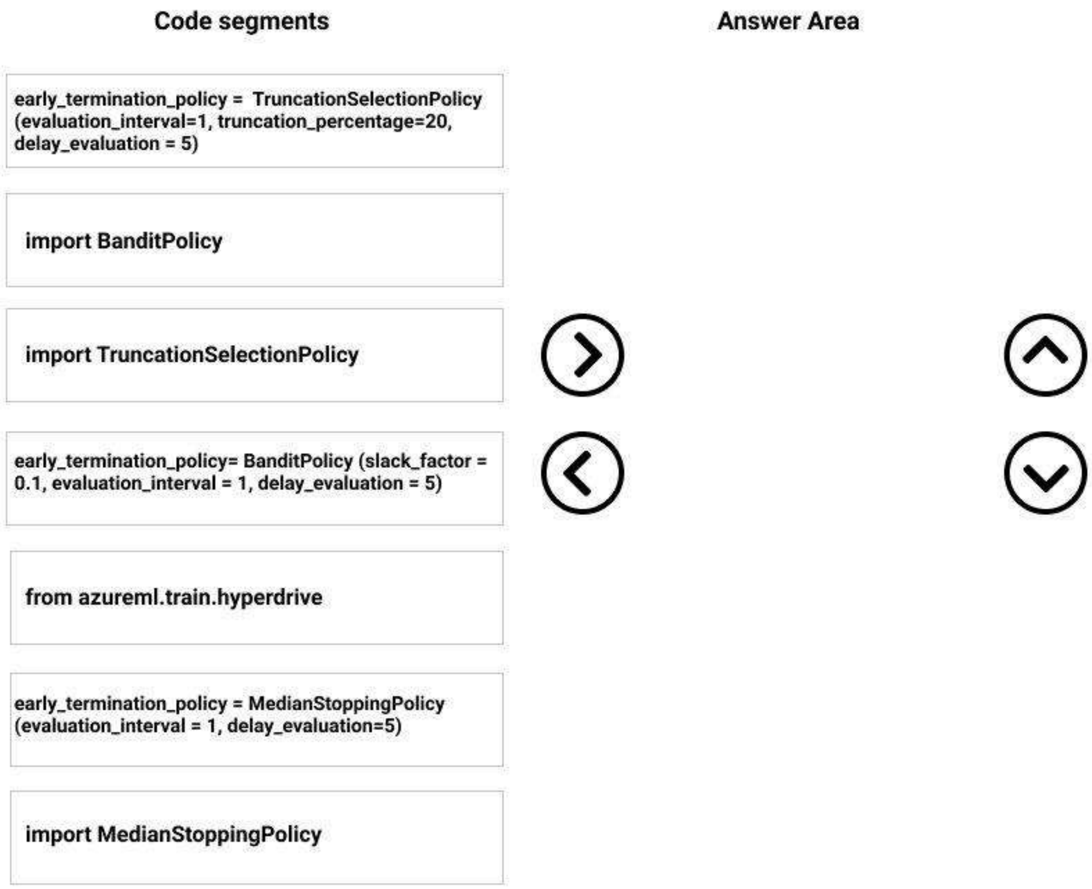
Testing -
You must produce multiple partitions of a dataset based on sampling using the Partition and Sample module in Azure Machine Learning Studio.

Cross-validation -
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. You must complete this task before the data goes through the sampling process.

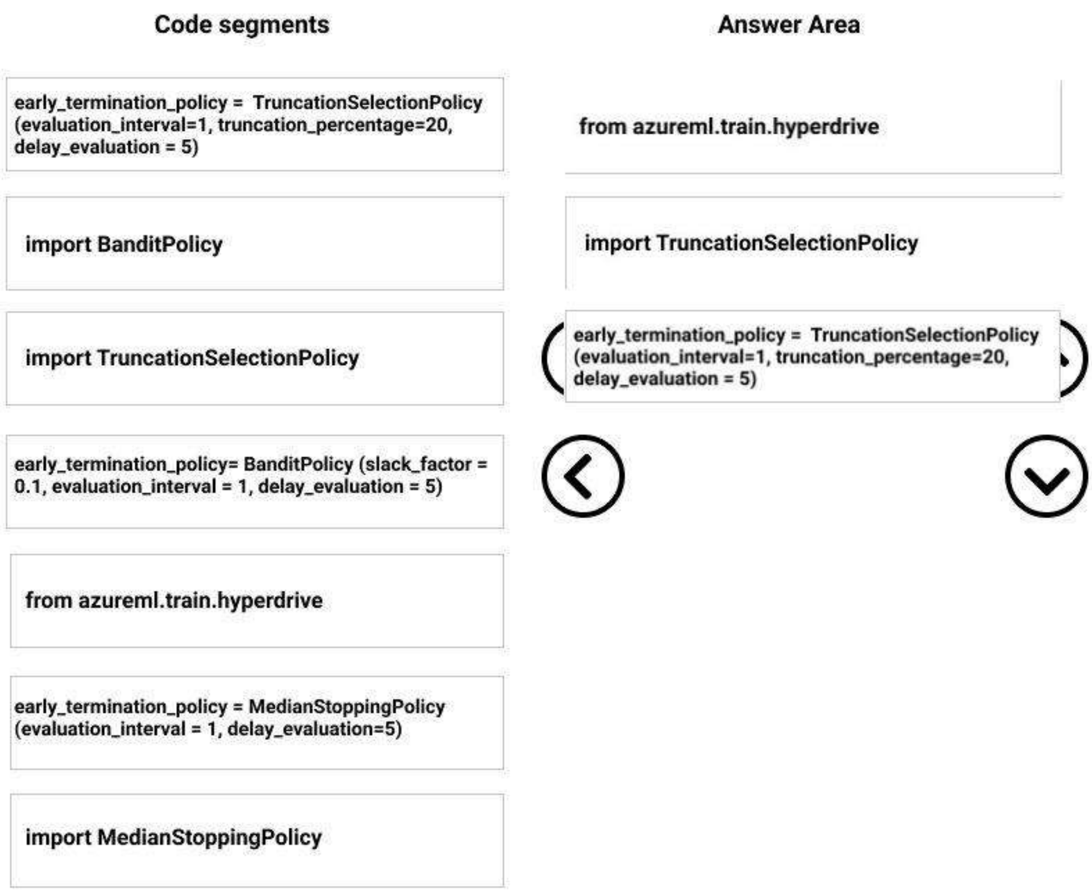
Linear regression module -
When you train a Linear Regression module, you must determine the best features to use in a model. You can choose standard metrics provided to measure performance before and after the feature importance process completes. The distribution of features across multiple training models must be consistent.

Data visualization -
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.**Question**DRAG DROP -
You need to implement early stopping criteria as stated 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 the credit for any of the correct orders you select.

Select and Place:



Correct Answer:



Step 1: from azureml.train.hyperdrive
Step 2: Import TruncationCelectionPolicy
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.
Scenario: You must configure hyperparameters in the model learning process to speed the learning phase. 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.
Step 3: early_terminiation_policy = TruncationSelectionPolicy..
Example:
from azureml.train.hyperdrive import TruncationSelectionPolicy early_termination_policy = TruncationSelectionPolicy(evaluation_interval=1, truncation_percentage=20, delay_evaluation=5)
In this example, the early termination policy is applied at every interval starting at evaluation interval 5. A run will be terminated at interval 5 if its performance at interval 5 is in the lowest 20% of performance of all runs at interval 5.
Incorrect Answers:
Median:
Median stopping is an early termination policy based on running averages of primary metrics reported by the runs. This policy computes running averages across all training runs and terminates runs whose performance is worse than the median of the running averages.
Slack:
Bandit is a termination policy based on slack factor/slack amount and evaluation interval. The policy early terminates any runs where the primary metric is not within the specified slack factor / slack amount with respect to the best performing training run.

Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/service/how-to-tune-hyperparameters

Topic 9 - Testlet 3
Question #1*Topic 9*

Introductory Info Case study -
This is a case study. Case studies are not timed separately. You can use as much exam time as you would like to complete each case. However, there may be additional case studies and sections on this exam. You must manage your time to ensure that you are able to complete all questions included on this exam in the time provided.
To answer the questions included in a case study, you will need to reference information that is provided in the case study. Case studies might contain exhibits and other resources that provide more information about the scenario that is described in the case study. Each question is independent of the other questions in this case study.
At the end of this case study, a review screen will appear. This screen allows you to review your answers and to make changes before you move to the next section of the exam. After you begin a new section, you cannot return to this section.

To start the case study -
To display the first question in this case study, click the Next button. Use the buttons in the left pane to explore the content of the case study before you answer the questions. Clicking these buttons displays information such as business requirements, existing environment, and problem statements. If the case study has an
All Information tab, note that the information displayed is identical to the information displayed on the subsequent tabs. When you are ready to answer a question, click the Question button to return to the question.

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. You add both files to Azure Machine Learning Studio as separate datasets to the starting point for an experiment. Both datasets contain 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 a 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

An initial investigation shows 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.

Data issues -

Missing values -
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 datasets also contain 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. 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-parametric statistics to measure relationships.
You must use a feature selection algorithm to analyze the relationship between the MedianValue and AvgRoomsInHouse columns.

Model training -

Permutation Feature Importance -
Given a trained model and a test dataset, you must compute the Permutation Feature Importance

scores of feature variables. You must be determined the absolute fit for the model.

Hyperparameters -
You must configure hyperparameters in the model learning process to speed the learning phase. 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, must 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.

Cross-validation -
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. You must complete this task before the data goes through the sampling process.

Linear regression module -
When you train a Linear Regression module, you must 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. The distribution of features across multiple training models must be consistent.

Data visualization -
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.**Question**DRAG DROP -
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.

Select and Place:

Actions

Answer Area

Add the Ordinal Regression module.

Add the Two-Class Averaged Perception module.

Augment the data.

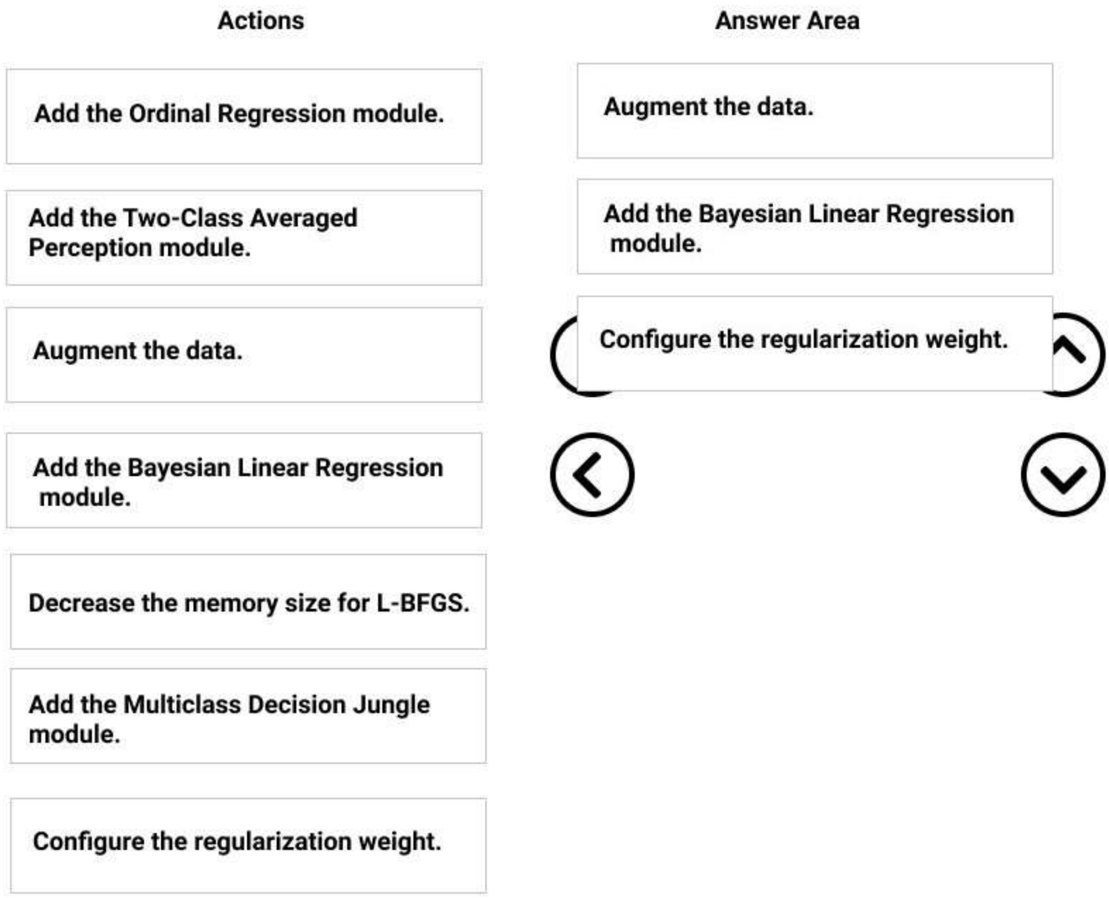
Add the Bayesian Linear Regression module.

Decrease the memory size for L-BFGS.

Add the Multiclass Decision Jungle module.

Configure the regularization weight.

Correct Answer:



Step 1: Augment the data -
Scenario: Columns in each dataset contain missing and null values. The datasets also contain many outliers.
Step 2: Add the Bayesian Linear Regression module.
Scenario: You produce a regression model to predict property prices by using the Linear Regression and Bayesian Linear Regression modules.
Step 3: Configure the regularization weight.
Regularization typically is used to avoid overfitting. For example, in L2 regularization weight, type the value to use as the weight for L2 regularization. We recommend that you use a non-zero value to avoid overfitting.
Scenario:
Model fit: The model shows signs of overfitting. You need to produce a more refined regression model that reduces the overfitting.
Incorrect Answers:
Multiclass Decision Jungle module:
Decision jungles are a recent extension to decision forests. A decision jungle consists of an ensemble of decision directed acyclic graphs (DAGs).
L-BFGS:
L-BFGS stands for "limited memory Broyden-Fletcher-Goldfarb-Shanno". It can be found in the wwo-Class Logistic Regression module, which is used to create a logistic regression model that can be used to predict two (and only two) outcomes.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/linear-regression