TEST 3 - QUESTION: 1/50

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 pandas as pd

data = pd.read_csv('./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('ml_data', data_folder,'data.csv'))
```


☐ C

```
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'))
```


☐ 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(data_folder,'train','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

KEEP OPEN

EXPLANATION:

Explanation: 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

TEST 3 - QUESTION: 2/50

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

Receiver Operating Characteristic (ROC) curve

D

Scatter plot

CORRECT ANSWER: C

KEEP OPEN

EXPLANATION:

Explanation: 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

TEST 3 - QUESTION: 3/50

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

KEEP OPEN

EXPLANATION:

Explanation: 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

TEST 3 - QUESTION: 4/50

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

KEEP OPEN

EXPLANATION:

Explanation: 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

TEST 3 - QUESTION: 5/50

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,l 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

No

☐ B

Yes

CORRECT ANSWER: A

KEEP OPEN

EXPLANATION:

Explanation: 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

TEST 3 - QUESTION: 6/50

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

Chi-squared

B

Spearman correlation

C

Kendall correlation

D

Pearson correlation

CORRECT ANSWER: D

KEEP OPEN

EXPLANATION:

Explanation: 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/

TEST 3 - QUESTION: 7/50

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.

CHECK BELOW THE RIGHT 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.	○	○

CORRECT ANSWER:

KEEP OPEN

EXPLANATION:

Explanation: 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

TEST 3 - QUESTION: 8/50

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

Execute Python Script

B

Filter Based Feature Selection

C

Edit Metadata

D

Latent Dirichlet Allocation

CORRECT ANSWER: C

KEEP OPEN

EXPLANATION:

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

TEST 3 - QUESTION: 9/50

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 batch_size

- [ ] B

a uniform expression for batch_size

- [ ] C

a uniform expression for learning_rate

- [ ] D

a normal expression for batch_size

- [ ] E

a choice expression for learning_rate

CORRECT ANSWERS: A,C

KEEP OPEN

EXPLANATION:

Explanation: 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

TEST 3 - QUESTION: 10/50

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.

CHECK BELOW THE RIGHT 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**

CORRECT ANSWER:

KEEP OPEN

EXPLANATION:

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

TEST 3 - QUESTION: 11/50

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.

CHECK BELOW THE RIGHT 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

CORRECT ANSWER:
KEEP OPEN

EXPLANATION:

Explanation: 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

TEST 3 - QUESTION: 12/50

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.

CHECK BELOW THE RIGHT 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

ORRECT ANSWER:

KEEP OPEN

EXPLANATION:

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

TEST 3 - QUESTION: 13/50

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=0.5

C

k=0.9

D

k=10

CORRECT ANSWER: D

KEEP OPEN

EXPLANATION:

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

TEST 3 - QUESTION: 14/50

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

No

☐ B

Yes

CORRECT ANSWER: A

KEEP OPEN

EXPLANATION:

Explanation: 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

TEST 3 - QUESTION: 15/50

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.log_row('row_count', rows)

☐ B

run.log('row_count', rows)

☐ C

run.tag('row_count', rows)

☐ D

run.log_table('row_count', rows)

☐ E

run.upload_file(T3 row_count', './data.csv')

CORRECT ANSWER: B

KEEP OPEN

EXPLANATION:

Explanation: 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

TEST 3 - QUESTION: 16/50

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.

CHECK BELOW THE RIGHT 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

CORRECT ANSWER:

KEEP OPEN

EXPLANATION:

Explanation: 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

TEST 3 - QUESTION: 17/50

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.

CHECK BELOW THE RIGHT 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()`	○	○

CORRECT ANSWER:

KEEP OPEN

EXPLANATION:

Explanation: 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

TEST 3 - QUESTION: 18/50

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 = list(automl_ex.get_runs())[0]
```

B

```
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]
```

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 = automl_ex.get_runs('AutoML_1234567890-123')
```

D

```
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
```

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.archived_time.find('11/11/2019 11:00:00 AM')
```

CORRECT ANSWER: B
KEEP OPEN

EXPLANATION:

Explanation: 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

TEST 3 - QUESTION: 19/50

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.

CHECK BELOW THE RIGHT ANSWER

Answer Area

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

CORRECT ANSWER:

KEEP OPEN

EXPLANATION:

Explanation: Box 1: PCA(n_components = 10) Need to reduce the dimensionality of the feature set to 10 features in both training and testing sets. Example: from sklearn.decomposition import PCA pca = PCA(n_components=2) ;2 dimensions principalComponents = pca.fit_transform(x) Box 2: pca fit_transform(X[, y]) fits the model with X and apply the dimensionality reduction on X. Box 3: transform(x_test) transform(X) applies dimensionality reduction to X. Reference: https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html

TEST 3 - QUESTION: 20/50

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.

CHECK BELOW THE RIGHT ANSWER

CORRECT ANSWER:

KEEP OPEN

EXPLANATION:

Explanation: 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

TEST 3 - QUESTION: 21/50

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_tf = True

C

exclude_nan_labels = True

D

enable_voting_ensemble = True

E

task = 'classification'

CORRECT ANSWER: A

KEEP OPEN

EXPLANATION:

Explanation: 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

TEST 3 - QUESTION: 22/50

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

No

B

Yes

CORRECT ANSWER: A

KEEP OPEN

EXPLANATION:

Explanation: 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

TEST 3 - QUESTION: 23/50

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 debug log

B
a file named parallel_run_step.txt located in the output folder

C
the digit_identification.py script

D
the Inference Clusters tab in Machine Learning studio

E
the Activity Log in the Azure portal for the Machine Learning workspace

CORRECT ANSWER: B

KEEP OPEN

EXPLANATION:

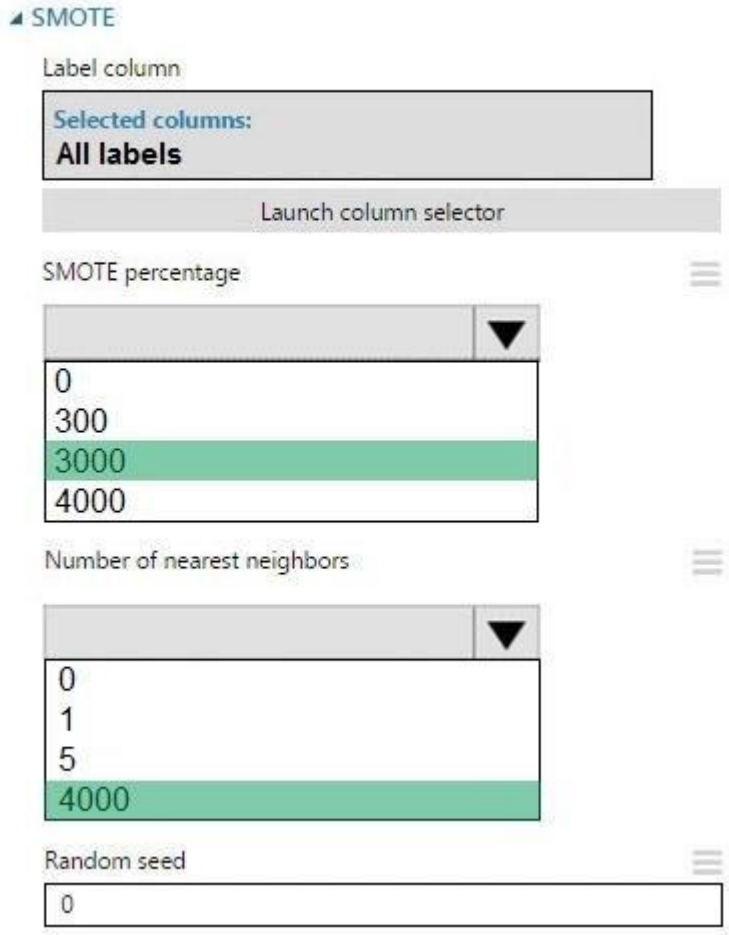
Explanation: output_action (str): How the output is to be organized. Currently supported values are 'append_row' and 'summary_only'. 'append_row' – All values output by run() method invocations will be aggregated into one unique file named parallel_run_step.txt that is created in the output location.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

TEST 3 - QUESTION: 24/50

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.

CHECK BELOW THE RIGHT ANSWER

Answer Area



CORRECT ANSWER:

KEEP OPEN

EXPLANATION:

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

TEST 3 - QUESTION: 25/50

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 accuracy of the model and log it as a float value with the key accuracy .

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 loss .

E

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 .

F

Set the primary_metric_name of the estimator used to run the bird_classifier_train.py script to accuracy .

CORRECT ANSWERS: A,B,F

KEEP OPEN

EXPLANATION:

Explanation: 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.

TEST 3 - QUESTION: 26/50

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

Create a config.yaml file defining the conda packages that are required and save the file in the experiment folder.

B

Create and register an Environment that includes the required packages. Use this Environment for all experiment runs.

C

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.

D

Always run the experiment with an Estimator by using the default packages.

E

Do not specify an environment in the run configuration for the experiment. Run the experiment by using the default environment.

CORRECT ANSWER: B

KEEP OPEN

EXPLANATION:

Explanation: 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

TEST 3 - QUESTION: 27/50

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.

CHECK BELOW THE RIGHT ANSWER

Answer Area

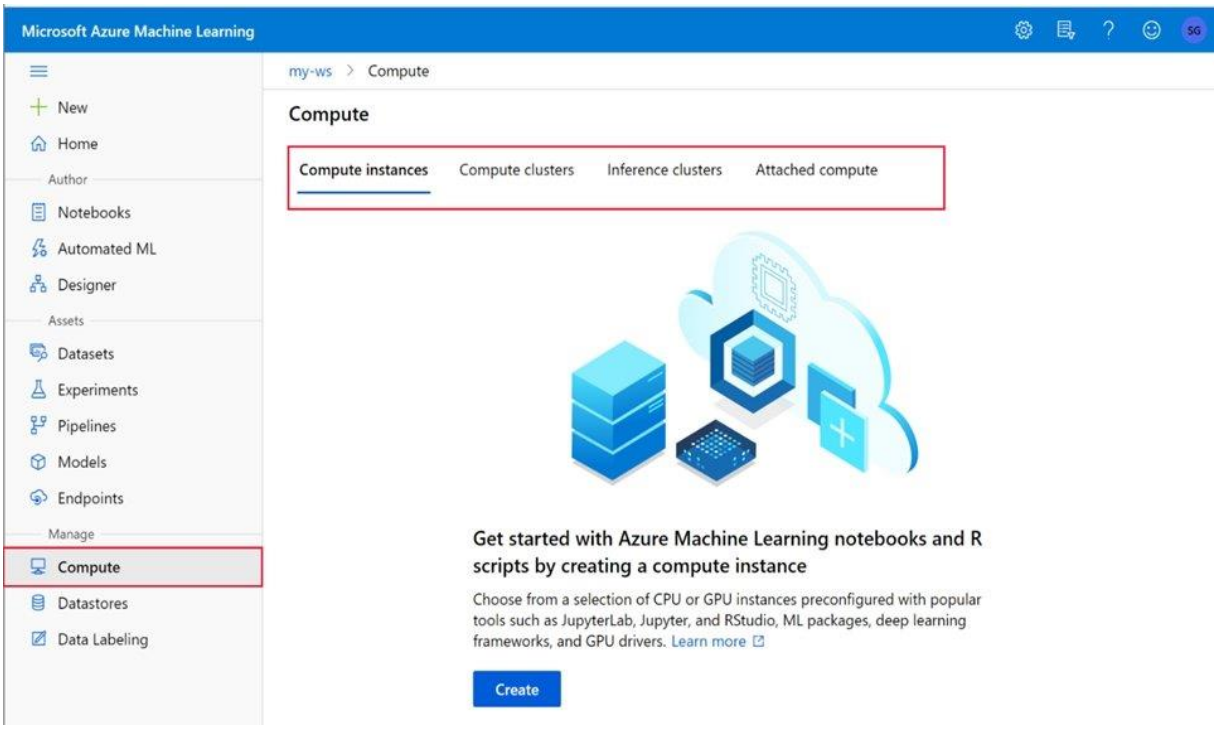
	Yes	No
The compute is created in the same region as the Machine Learning service workspace.	○	○
The compute resource created by the code is displayed as a compute cluster in Azure Machine Learning studio.	○	○
The minimum number of nodes will be zero.	○	○

CORRECT ANSWER:

KEEP OPEN

EXPLANATION:

Explanation: 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

TEST 3 - QUESTION: 28/50

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

Gradient descent

B

Box plot

C

Violin plot

D

Binary classification confusion matrix

CORRECT ANSWER: D

KEEP OPEN

EXPLANATION:

Explanation: 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/

TEST 3 - QUESTION: 29/50

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.

CHECK BELOW THE RIGHT ANSWER

Answer Area

```
import pandas as pd
salesData = pd.melt(
```

[▼] ,id_vars=' [▼] ',value_vars= [▼])

dataFrame	shop	'shop'
pandas	year	'year'
salesData	value	['year']
year	Shop X, Shop Y, Shop Z	['2017', '2018]

CORRECT ANSWER:

KEEP OPEN

EXPLANATION:

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

TEST 3 - QUESTION: 30/50

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

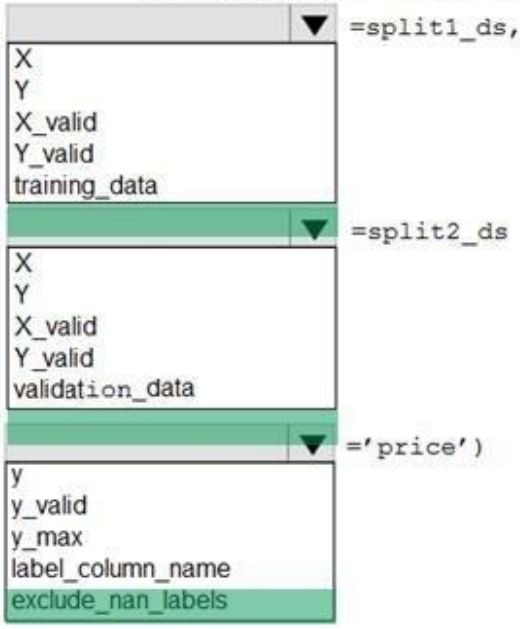
You want to use automated machine learning to find the best regression model for predicting the price column. You need to configure an automated machine learning experiment using the Azure Machine Learning SDK. How should you complete the code? To answer, select the appropriate options in the answer area. NOTE: Each correct selection is worth one point.

CHECK BELOW THE RIGHT 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',
```



CORRECT ANSWER:
KEEP OPEN

EXPLANATION:

Explanation: 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

TEST 3 - QUESTION: 31/50

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 non-discrete.

☐ C

The label data can be positive or negative.

☐ D

The label data must be whole numbers.

☐ E

The label data must be a positive value.

CORRECT ANSWERS: D,E

KEEP OPEN

EXPLANATION:

Explanation: Poisson regression is intended for use in regression models that are used to predict numeric values, typically counts. Therefore, you should use this module to create your regression model only if the values you are trying to predict fit the following conditions: The response variable has a Poisson distribution. Counts cannot be negative. The method will fail outright if you attempt to use it with negative labels. A Poisson distribution is a discrete distribution; therefore, it is not meaningful to use this method with non-whole numbers. Reference: https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/poisson-regression

TEST 3 - QUESTION: 32/50

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

KEEP OPEN

EXPLANATION:

Explanation: 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

TEST 3 - QUESTION: 33/50

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

ROC curve

B

Random forest diagram

C

Box plot

D

Venn diagram

E

Scatter plot

CORRECT ANSWERS: C,E

KEEP OPEN

EXPLANATION:

Explanation: 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/

TEST 3 - QUESTION: 34/50

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,l 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: B

KEEP OPEN

EXPLANATION:

Explanation: 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

TEST 3 - QUESTION: 35/50

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

Set the task type to Regression .

B

Clear the option to enable deep learning.

C

Clear the option to perform automatic featurization.

D

Add all algorithms other than linear ones to the blocked algorithms list.

E

Set the Exit criterion option to a metric score threshold.

CORRECT ANSWER: C

KEEP OPEN

EXPLANATION:

Explanation: 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

TEST 3 - QUESTION: 36/50

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.

CHECK BELOW THE RIGHT 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.	○	○

CORRECT ANSWER:

KEEP OPEN

EXPLANATION:

Explanation: 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.

TEST 3 - QUESTION: 37/50

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: D

KEEP OPEN

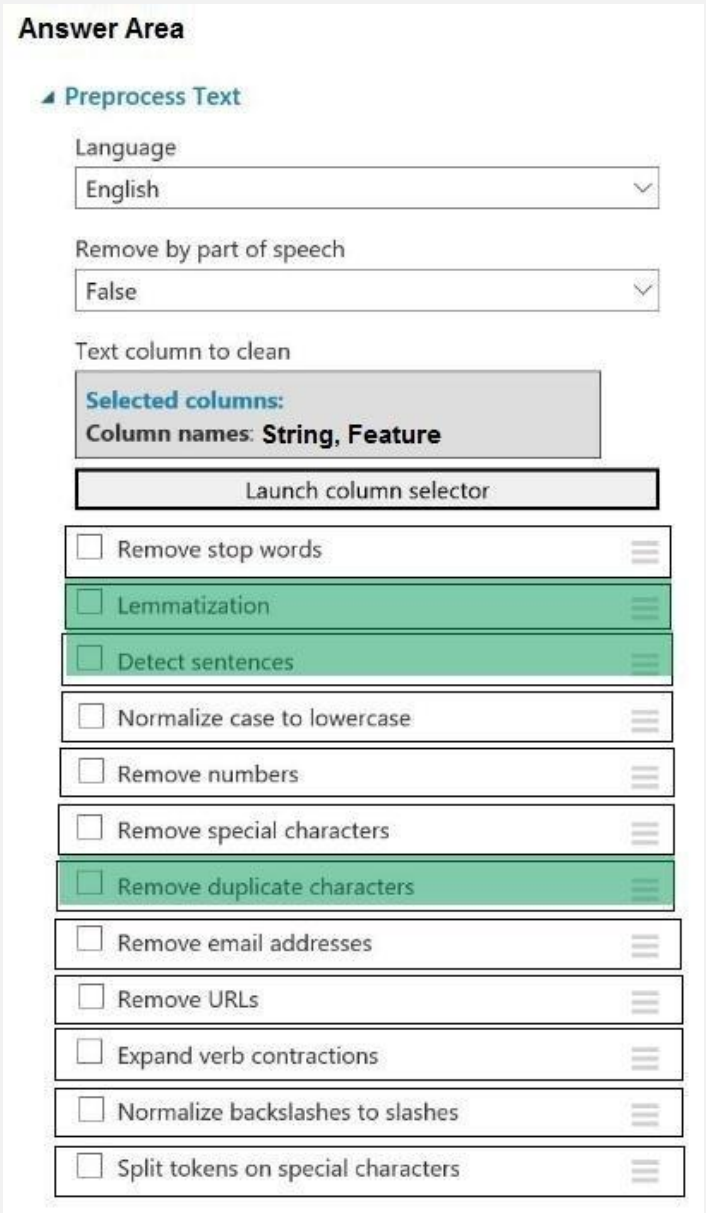
EXPLANATION:

Explanation: 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

TEST 3 - QUESTION: 38/50

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.

CHECK BELOW THE RIGHT ANSWER



CORRECT ANSWER:

KEEP OPEN

EXPLANATION:

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

TEST 3 - QUESTION: 39/50

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

No

B

Yes

CORRECT ANSWER: A

KEEP OPEN

EXPLANATION:

Explanation: 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

TEST 3 - QUESTION: 40/50

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.

CHECK BELOW THE RIGHT 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.	○	○

CORRECT ANSWER:

KEEP OPEN

EXPLANATION:

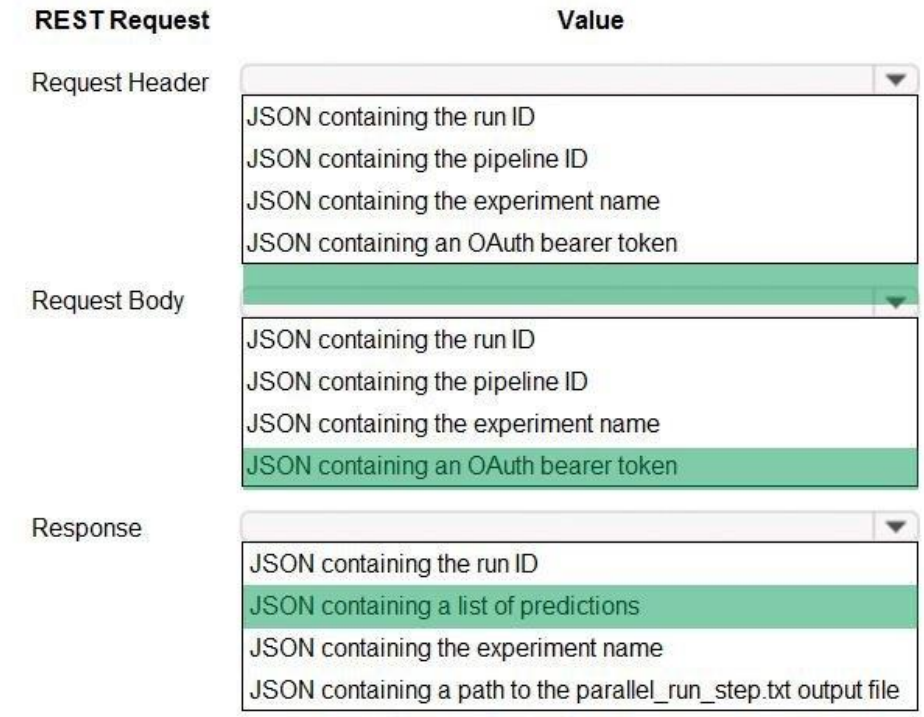
Explanation: Box 1: No max_total_runs (50 here) The maximum total number of runs to create. This is the upper bound; there may be fewer runs when the sample space is smaller than this value. Box 2: Yes Policy EarlyTerminationPolicy The early termination policy to use. If None - the default, no early termination policy will be used. Box 3: No Discrete hyperparameters are specified as a choice among discrete values.

TEST 3 - QUESTION: 42/50

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.

CHECK BELOW THE RIGHT ANSWER

Answer Area



CORRECT ANSWER:

KEEP OPEN

EXPLANATION:

Explanation: 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

TEST 3 - QUESTION: 43/50

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

source_directory = training_ds

B

environment_definition = {"training_data":training_ds}

C

script_params = {"--training_ds":training_ds}

D

inputs = [training_ds.as_named_input('training_ds')]

CORRECT ANSWER: D

KEEP OPEN

EXPLANATION:

Explanation: 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

TEST 3 - QUESTION: 44/50

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

No

☐ B

Yes

CORRECT ANSWER: A

KEEP OPEN

EXPLANATION:

Explanation: 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

TEST 3 - QUESTION: 45/50

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.

CHECK BELOW THE RIGHT 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.	○	○

CORRECT ANSWER:

KEEP OPEN

EXPLANATION:

Explanation: 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

TEST 3 - QUESTION: 46/50

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').to_pandas_dataframe()],

B
script_params={'--training_files': file_dataset},

C
inputs=[file_dataset.as_named_input('training_files').as_mount()],

D
inputs=[file_dataset.as_named_input('training_files')],

CORRECT ANSWER: C

KEEP OPEN

EXPLANATION:

Explanation: 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

TEST 3 - QUESTION: 47/50

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,l 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

No

B

Yes

CORRECT ANSWER: B

KEEP OPEN

EXPLANATION:

Explanation: 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:

TEST 3 - QUESTION: 48/50

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=5

C

k=0.01

D

k=0.5

CORRECT ANSWER: B

KEEP OPEN

EXPLANATION:

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

TEST 3 - QUESTION: 49/50

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.

CHECK BELOW THE RIGHT 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
    [ run.log()
      mlflow.log_metric
      print ] ('my_metric', 1.00)
print("Finished!")
```

CORRECT ANSWER:

KEEP OPEN

EXPLANATION:

Explanation: 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

TEST 3 - QUESTION: 50/50

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: A

KEEP OPEN

EXPLANATION:

Explanation: 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