DUMPSARENA

Designing and Implementing a Data Science Solution on Azure

Microsoft DP-100

Version Demo

Total Demo Questions: 20

Total Premium Questions: 502

Buy Premium PDF

https://dumpsarena.co

sales@dumpsarena.co

sales@dumpsarena.co
dumpsarena.co

Topic Break Down

Topic	No. of Questions
Topic 2, New Update	208
Topic 3, Case Study 1	8
Topic 4, Case Study 2	12
Topic 5, Mixed Questions	274
Total	502

QUESTION NO: 1

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

ANSWER: A C

Explanation:

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 NO: 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 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

ANSWER: B

Explanation:

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 NO: 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

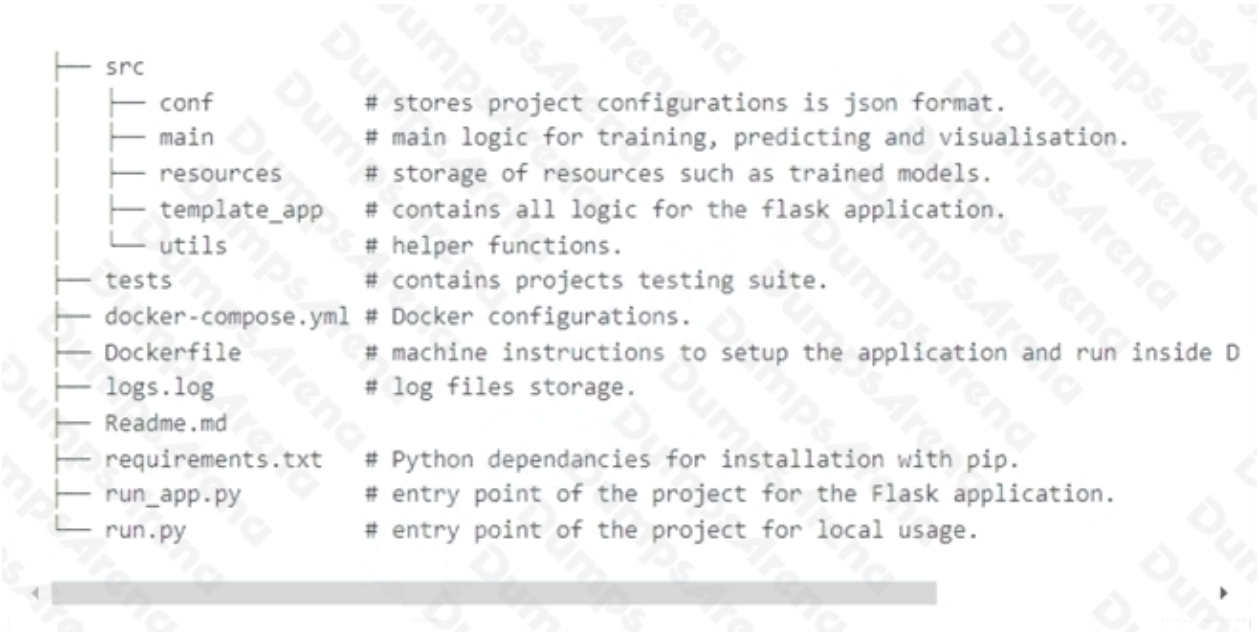
ANSWER: A C

Explanation:

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.
```

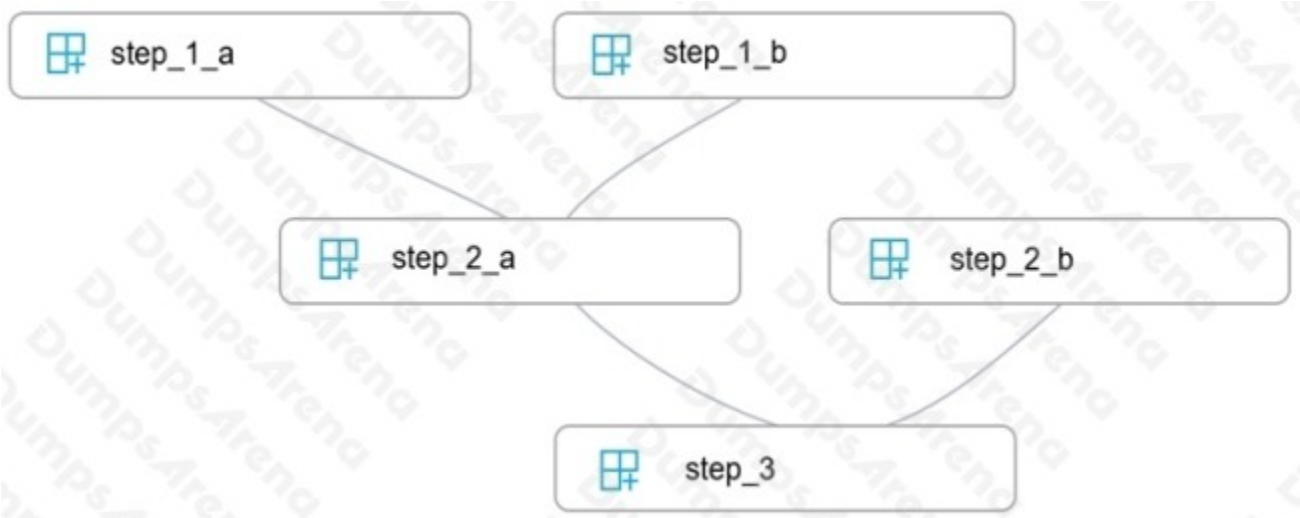
Once the project is nicely structured we can tidy up or refactor the code.

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 NO: 4

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])
```

A. Option A

B. Option B

C. Option C

D. Option D

ANSWER: A

Explanation:

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 NO: 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

ANSWER: A B D

Explanation:

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 NO: 6

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

ANSWER: C

Explanation:

Leave One Out (LOO) cross-validation

Setting K = n (the number of observations) yields n-fold and is called leave-one out cross-validation (LOO), a special case of the K-fold approach.

LOO CV is sometimes useful but typically doesn't shake up the data enough. The estimates from each fold are highly correlated and hence their average can have high variance. This is why the usual choice is K=5 or 10. It provides a good compromise for the bias-variance tradeoff.

QUESTION NO: 7

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. Local web service

B. Remote VM

C. Azure Databricks

D. Azure Machine Learning Kubernetes

E. Azure Container Instances

ANSWER: A E

QUESTION NO: 8

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.

D

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.

ANSWER: A B D

Explanation:

Reference:

https://azure.microsoft.com/en-us/pricing/details/machine-learning/

QUESTION NO: 9

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

Subscription	Workspace	Comment
385bdfe5-4cef-4ad4-b977-3f86d92727c9	ml-default	This is the 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

ANSWER: B

QUESTION NO: 10

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

ANSWER: D

Explanation:

Maximum number of runs on random grid: This option also controls the number of iterations over a random sampling of parameter values, but the values are not generated randomly from the specified range; instead, a matrix is created of all possible combinations of parameter values and a random sampling is taken over the matrix. This method is more efficient and less prone to regional oversampling or undersampling.

If you are training a model that supports an integrated parameter sweep, you can also set a range of seed values to use and iterate over the random seeds as well. This is optional, but can be useful for avoiding bias introduced by seed selection.

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 NO: 11 - (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 panty constraint should you use? To answer, select the appropriate options in the answer area. NOTE: Each correct selection is worth one point.

Answer Area

ANSWER:

Answer Area

Explanation:

QUESTION NO: 12

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

ANSWER: A D

Explanation:

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 underprediction.

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 NO: 13

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)

ANSWER: D

Explanation:

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

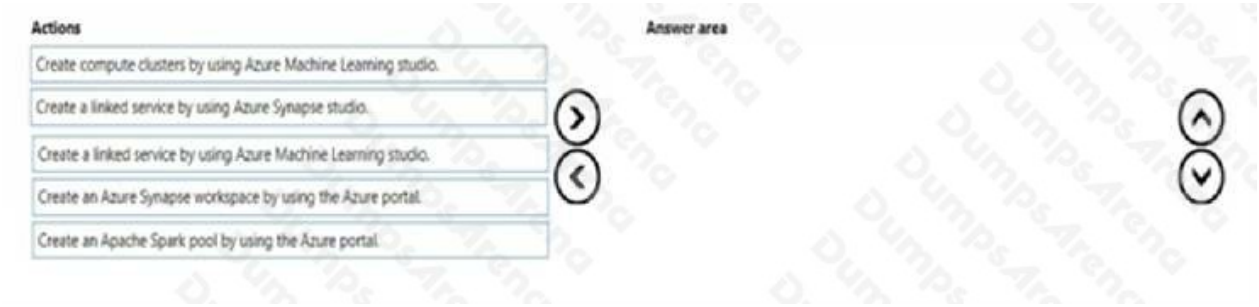
QUESTION NO: 14 - (DRAG DROP)

You create an Azure Machine Learning workspace.

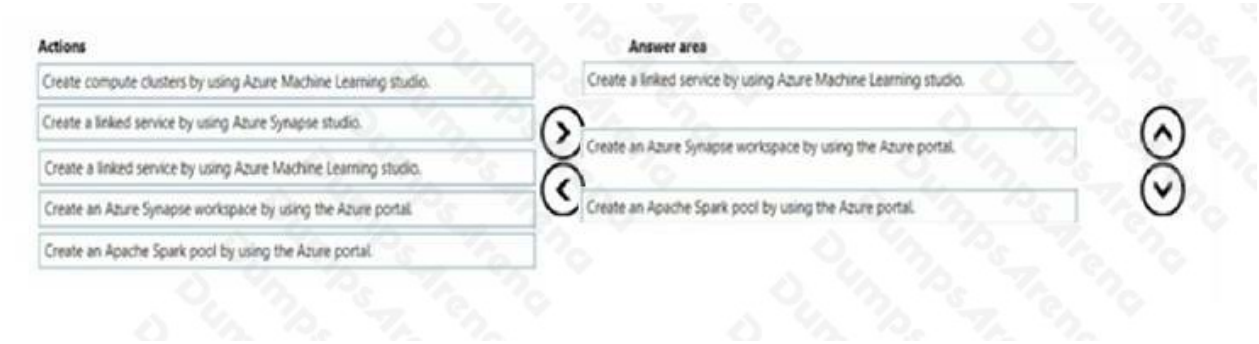
You must implement dedicated compute for model training in the workspace by using Azure Synapse compute resources. The solution must attach the dedicated compute and start an Azure Synapse session.

You need to implement the compute resources.

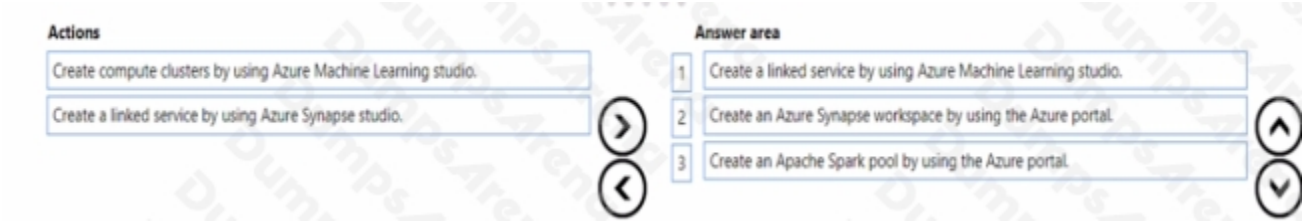
Which three actions should you perform in sequence? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.



ANSWER:



Explanation:



QUESTION NO: 15

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.

ANSWER: B C

Explanation:

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 NO: 16

You manage an Azure Machine Learning workspace named workspaces

You must develop Python SDK v2 code to attach an Azure Synapse Spark pool as a compute target in workspaces 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

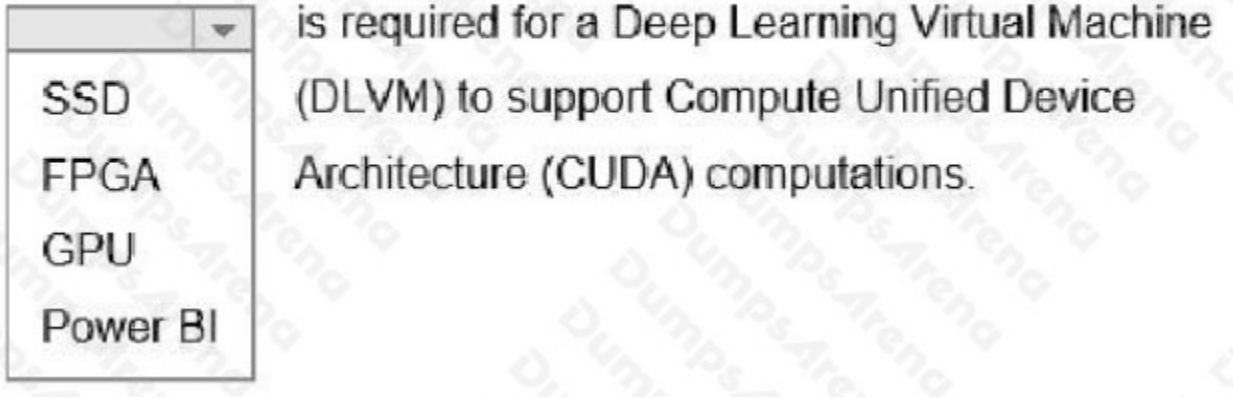
ANSWER: B

QUESTION NO: 17 - (HOTSPOT)

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

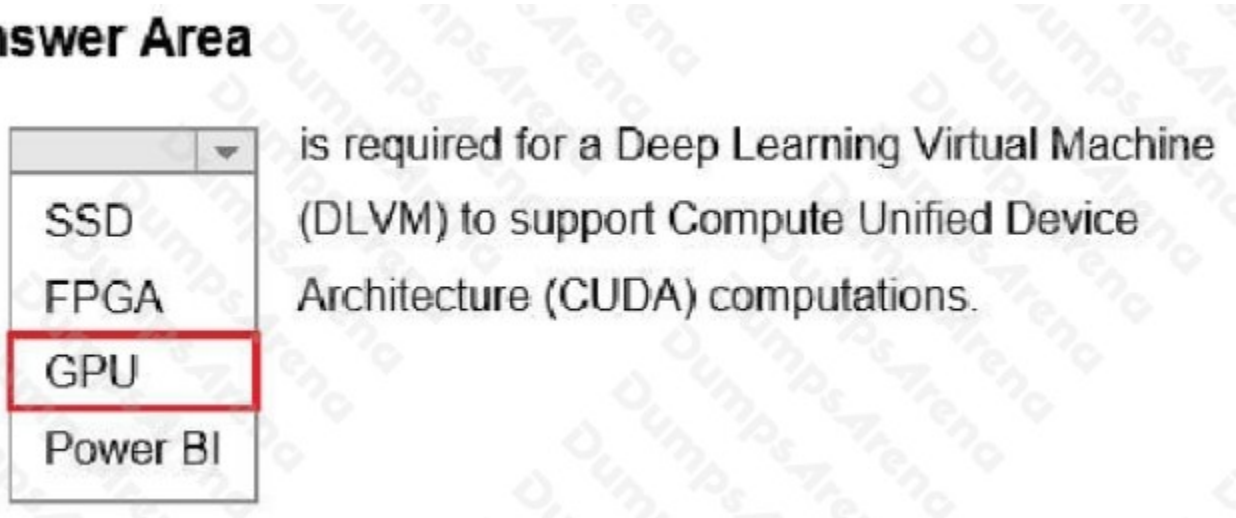
Hot Area:

Answer Area



ANSWER:

Answer Area



Explanation:

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

References: https://azuremarketplace.microsoft.com/en-au/marketplace/apps/microsoft-ads.dsvm-deep-learning

QUESTION NO: 18 - (SIMULATION)

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 range 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.

ANSWER: seetheanswerbelow.

Explanation:

See below image

QUESTION NO: 19 - (SIMULATION)

You create an Azure Data Lake Storage Gen2 stowage account named storage1 containing a file system named fsi 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.

ANSWER: seetheanswerbelow.

Explanation:

See below image

QUESTION NO: 20 - (DRAG DROP)

DRAG DROP

You are building an intelligent solution using machine learning models.

The environment must support the following requirements:

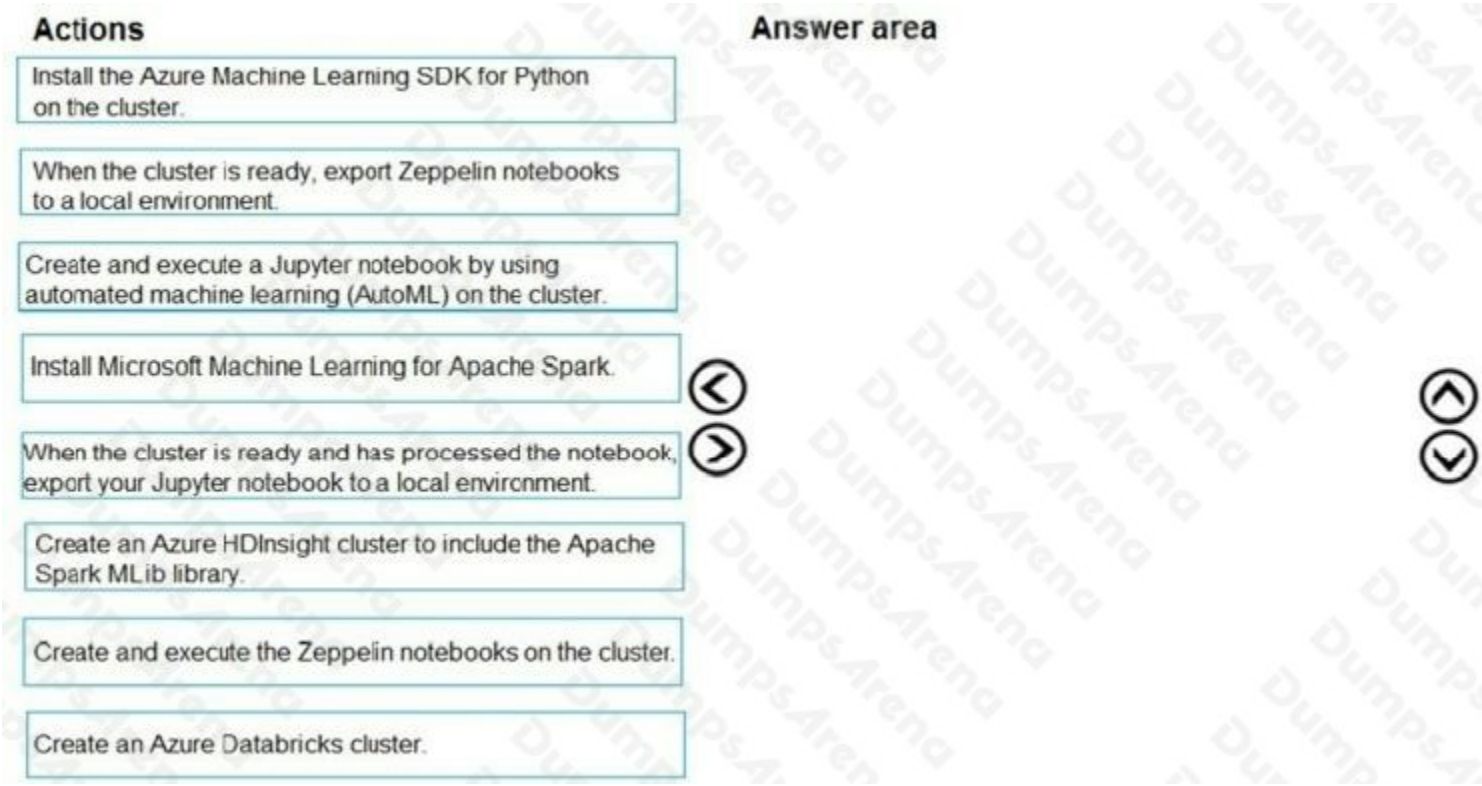
- Data scientists must build notebooks in a cloud environment
- Data scientists must use automatic feature engineering and model building in machine learning pipelines.

▪ Notebooks must be deployed to retrain using Spark instances with dynamic worker allocation. ▪ Notebooks must be exportable to be version controlled locally.

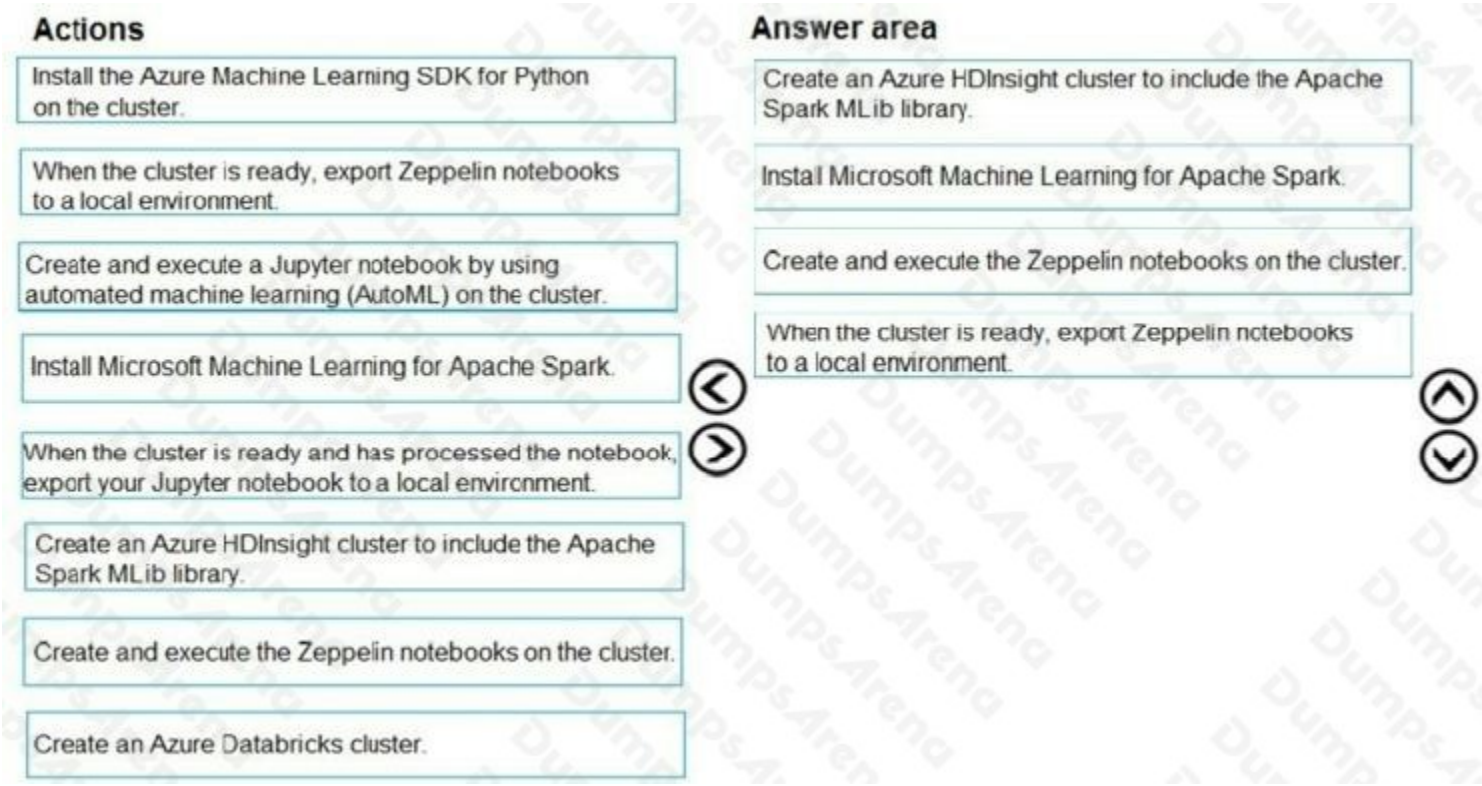
You need to create the environment.

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:



ANSWER:



Explanation:

Step 1: Create an Azure HDInsight cluster to include the Apache Spark Mlib library

Step 2: Install Microsot Machine Learning for Apache Spark You install AzureML on your Azure HDInsight cluster.

Microsoft Machine Learning for Apache Spark (MMLSpark) provides a number of deep learning and data science tools for Apache Spark, including seamless integration of Spark Machine Learning pipelines with Microsoft Cognitive Toolkit (CNTK) and OpenCV, enabling you to quickly create powerful, highly-scalable predictive and analytical models for large image and text datasets.

Step 3: Create and execute the Zeppelin notebooks on the cluster

Step 4: When the cluster is ready, export Zeppelin notebooks to a local environment. Notebooks must be exportable to be version controlled locally.

References: https://docs.microsoft.com/en-us/azure/hdinsight/spark/apache-spark-zeppelin-notebook https://azuremlbuild.blob.core.windows.net/pysparkapi/intro.html