DP-100 Dumps

Designing and Implementing a Data Science Solution on Azure

https://www.certleader.com/DP-100-dumps.html

NEW QUESTION 1
- (Exam Topic 3)
You are determining if two sets of data are significantly different from one another by using Azure Machine Learning Studio.
Estimated values in one set of data may be more than or less than reference values in the other set of data. You must produce a distribution that has a constant Type I error as a function of the correlation.
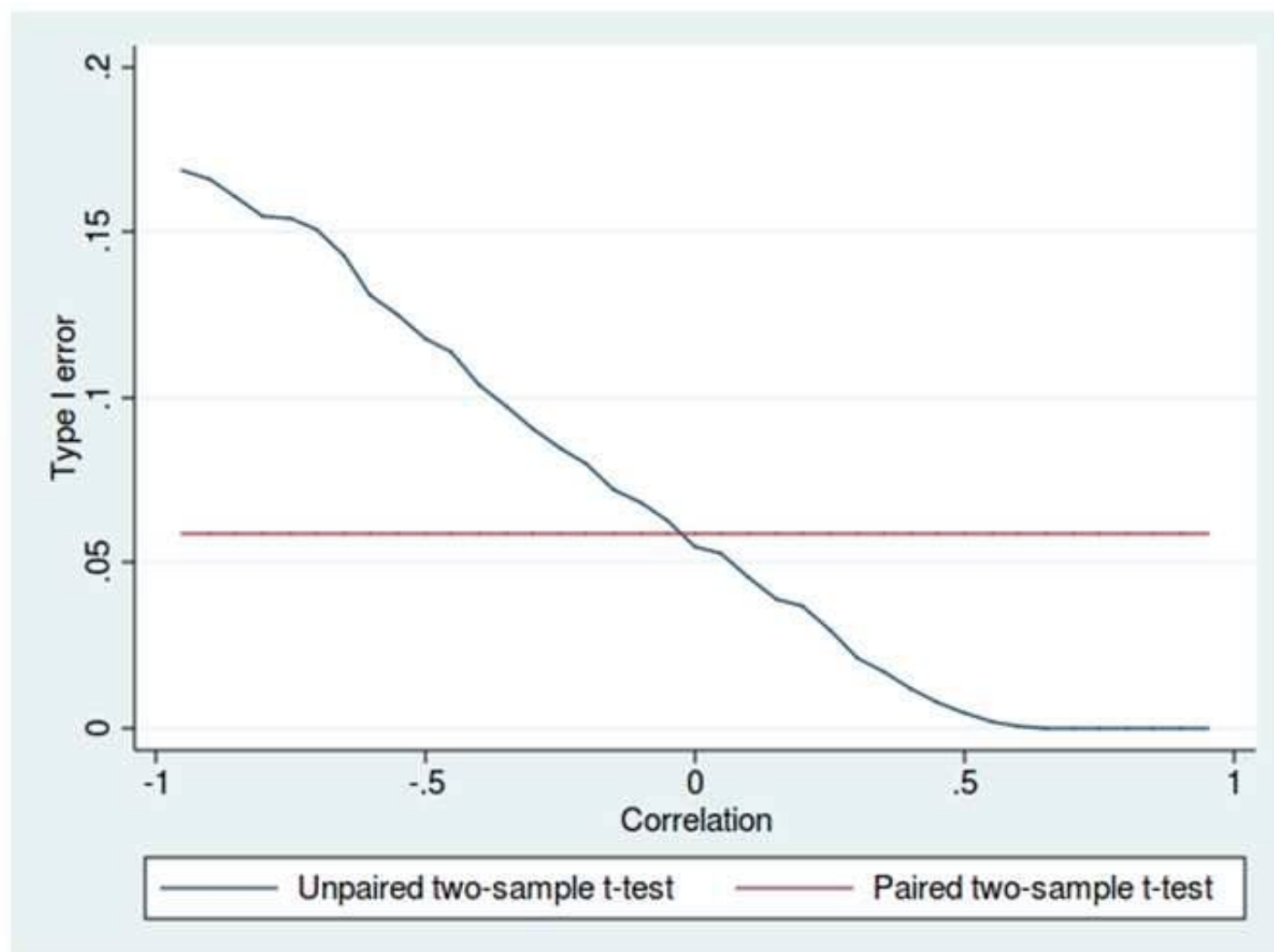You need to produce the distribution.
Which type of distribution should you produce?

A. Paired t-test with a two-tail option
B. Unpaired t-test with a two tail option
C. Paired t-test with a one-tail option
D. Unpaired t-test with a one-tail option

Answer: A

Explanation:
Choose a one-tail or two-tail test. The default is a two-tailed test. This is the most common type of test, in which the expected distribution is symmetric around zero.
Example: Type I error of unpaired and paired two-sample t-tests as a function of the correlation. The simulated random numbers originate from a bivariate normal distribution with a variance of 1.

Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/test-hypothesis-using-t-test https://en.wikipedia.org/wiki/Student%27s_t-test

NEW QUESTION 2
- (Exam Topic 3)
You plan to deliver a hands-on workshop to several students. The workshop will focus on creating data visualizations using Python. Each student will use a device that has internet access.
Student devices are not configured for Python development. Students do not have administrator access to install software on their devices. Azure subscriptions are not available for students.
You need to ensure that students can run Python-based data visualization code. Which Azure tool should you use?

A. Anaconda Data Science Platform
B. Azure BatchAI
C. Azure Notebooks
D. Azure Machine Learning Service

Answer: C

Explanation:
References: https://notebooks.azure.com/

NEW QUESTION 3
- (Exam Topic 3)
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.
You are creating a model to predict the price of a student's artwork depending on the following variables: the student's length of education, degree type, and art form.
You start by creating a linear regression model. You need to evaluate the linear regression model.

Solution: Use the following metrics: Mean Absolute Error, Root Mean Absolute Error, Relative Absolute Error, Relative Squared Error, and the Coefficient of Determination.
Does the solution meet the goal?

A. Yes
B. No

Answer: A

Explanation:
The following metrics are reported for evaluating regression models. When you compare models, they are ranked by the metric you select for evaluation.
Mean absolute error (MAE) measures how close the predictions are to the actual outcomes; thus, a lower score is better.
Root mean squared error (RMSE) creates a single value that summarizes the error in the model. By squaring the difference, the metric disregards the difference between over-prediction and under-prediction.
Relative absolute error (RAE) is the relative absolute difference between expected and actual values; relative because the mean difference is divided by the arithmetic mean.
Relative squared error (RSE) similarly normalizes the total squared error of the predicted values by dividing by the total squared error of the actual values.
Mean Zero One Error (MZOE) indicates whether the prediction was correct or not. In other words: ZeroOneLoss(x,y) = 1 when x!=y; otherwise 0.
Coefficient of determination, often referred to as R2, represents the predictive power of the model as a value between 0 and 1. Zero means the model is random (explains nothing); 1 means there is a perfect fit. However, caution should be used in interpreting R2 values, as low values can be entirely normal and high values can be suspect.
AUC.
References:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/evaluate-model

NEW QUESTION 4
- (Exam Topic 3)
You are moving a large dataset from Azure Machine Learning Studio to a Weka environment. You need to format the data for the Weka environment.
Which module should you use?

A. Convert to CSV
B. Convert to Dataset
C. Convert to ARFF
D. Convert to SVMLight

Answer: C

Explanation:
Use the Convert to ARFF module in Azure Machine Learning Studio, to convert datasets and results in Azure Machine Learning to the attribute-relation file format used by the Weka toolset. This format is known as ARFF.
The ARFF data specification for Weka supports multiple machine learning tasks, including data preprocessing, classification, and feature selection. In this format, data is organized by entites and their attributes, and is contained in a single text file.
References:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/convert-to-arff

NEW QUESTION 5
- (Exam Topic 3)
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.

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

A. Mastered
B. Not Mastered

Answer: A

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-2d2b8565

NEW QUESTION 6
- (Exam Topic 3)
You arc creating a new experiment in Azure Machine Learning Studio. You have a small dataset that has missing values in many columns. The data does not require the application of predictors for each column. You plan to use the Clean Missing Data module to handle the missing data.
You need to select a data cleaning method. Which method should you use?

A. Synthetic Minority
B. Replace using Probabilistic PAC
C. Replace using MICE
D. Normalization

Answer: B

NEW QUESTION 7
- (Exam Topic 3)
You are analyzing a dataset by using Azure Machine Learning Studio.
YOU need to generate a statistical summary that contains the p value and the unique value count for each feature column.
Which two modules can you users? Each correct answer presents a complete solution. NOTE: Each correct selection is worth one point.

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

Answer: BE

Explanation:
The Export Count Table module is provided for backward compatibility with experiments that use the Build Count Table (deprecated) and Count Featurizer (deprecated) modules.
E: Summarize Data statistics are useful when you want to understand the characteristics of the complete dataset. For example, you might need to know:
How many missing values are there in each column? How many unique values are there in a feature column?
What is the mean and standard deviation for each column?
The module calculates the important scores for each column, and returns a row of summary statistics for each variable (data column) provided as input.
References:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/export-count-table https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/summarize-data

NEW QUESTION 8

- (Exam Topic 3)
You are with a time series dataset in Azure Machine Learning Studio.
You need to split your dataset into training and testing subsets by using the Split Data module. Which splitting mode should you use?

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

Answer: B

Explanation:
Split Rows: Use this option if you just want to divide the data into two parts. You can specify the percentage of data to put in each split, but by default, the data is divided 50-50.
References:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/split-data

NEW QUESTION 9
- (Exam Topic 3)
You are analyzing a dataset containing historical data from a local taxi company. You arc developing a regression a regression model.
You must predict the fare of a taxi trip.
You need to select performance metrics to correctly evaluate the- regression model. Which two metrics can you use? Each correct answer presents a complete solution. NOTE: Each correct selection is worth one point.

A. an F1 score that is high
B. an R Squared value dose to 1
C. an R-Squared value close to 0
D. a Root Mean Square Error value that is high
E. a Root Mean Square Error value that is tow
F. an F 1 score that is low.

Answer: DF

NEW QUESTION 10
- (Exam Topic 3)
You have a dataset that contains over 150 features. You use the dataset to train a Support Vector Machine (SVM) binary classifier.
You need to use the Permutation Feature Importance module in Azure Machine Learning Studio to compute a set of feature importance scores for the dataset.
In which order should you perform the actions? To answer, move all actions from the list of actions to the answer area and arrange them in the correct order.

Actions	Answer Area
Add a Two-Class Support Vector Machine module to initialize the SVM classifier.	
Set the Metric for measuring performance property to **Classification - Accuracy** and then run the experiment.	
Add a Permutation Feature Importance module and connect the trained model and test dataset.	
Add a dataset to the experiment.	
Add a Split Data module to create training and test datasets.	

A. Mastered
B. Not Mastered

Answer: A

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

NEW QUESTION 10
- (Exam Topic 3)
You are a data scientist building a deep convolutional neural network (CNN) for image classification. The CNN model you built shows signs of overfitting.
You need to reduce overfitting and converge the model to an optimal fit.
Which two actions should you perform? Each correct answer presents a complete solution. NOTE: Each correct selection is worth one point.

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

Answer: AC

Explanation:
 References:
https://machinelearningmastery.com/how-to-reduce-overfitting-in-deep-learning-with-weight-regularization/
https://en.wikipedia.org/wiki/Convolutional_neural_network

NEW QUESTION 13
- (Exam Topic 2)
You need to select a feature extraction method. Which method should you use?

A. Mutual information
B. Mood's median test
C. Kendall correlation
D. Permutation Feature Importance

Answer: C

Explanation:
In statistics, the Kendall rank correlation coefficient, commonly referred to as Kendall's tau coefficient (after the Greek letter), is a statistic used to measure the ordinal association between two measured quantities.
It is a supported method of the Azure Machine Learning Feature selection.
Scenario: When you train a Linear Regression module using a property dataset that shows data for property prices for a large city, you need to determine the best features to use in a model. You can choose standard metrics provided to measure performance before and after the feature importance process completes. You must ensure that the distribution of the features across multiple training models is consistent.
References:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/feature-selection-modules

NEW QUESTION 16
- (Exam Topic 2)
You need to visually identify whether outliers exist in the Age column and quantify the outliers before the outliers are removed.
Which three Azure Machine Learning Studio modules should you use in sequence? To answer, move the appropriate modules from the list of modules to the answer area and arrange them in the correct order.
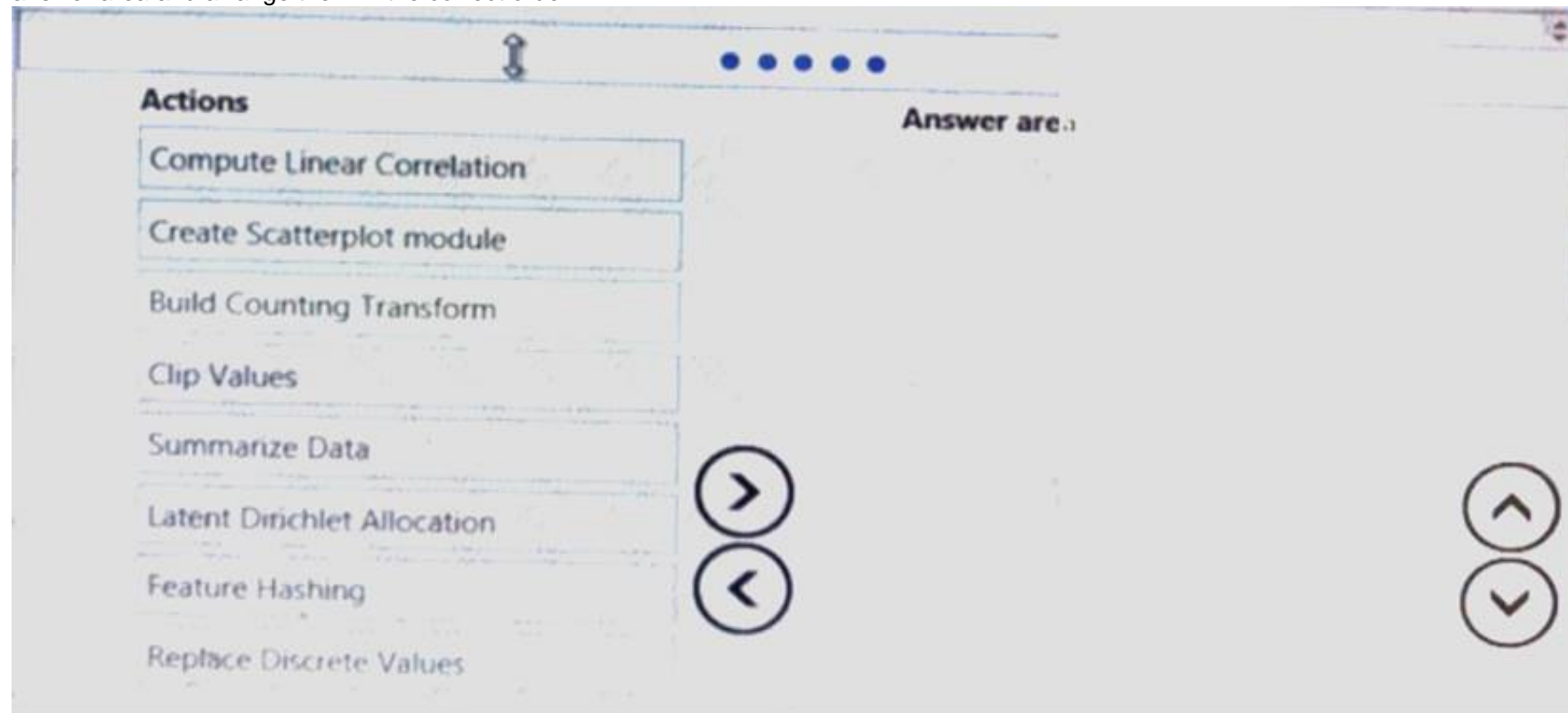
A. Mastered
B. Not Mastered

Answer: A

Explanation:
 Create Scatterplot Summarize Data Clip Values
You can use the Clip Values module in Azure Machine Learning Studio, to identify and optionally replace data values that are above or below a specified threshold. This is useful when you want to remove outliers or replace them with a mean, a constant, or other substitute value.
References:
https://blogs.msdn.microsoft.com/azuredev/2017/05/27/data-cleansing-tools-in-azure-machine-learning/ https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/clip-values

NEW QUESTION 19
- (Exam Topic 3)
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the

stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.

After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.

You are a data scientist using Azure Machine Learning Studio.

You need to normalize values to produce an output column into bins to predict a target column. Solution: Apply an Equal Width with Custom Start and Stop binning mode.

Does the solution meet the goal?

A. Yes
B. No

Answer: B

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

NEW QUESTION 22
- (Exam Topic 2)
You need to correct the model fit issue.

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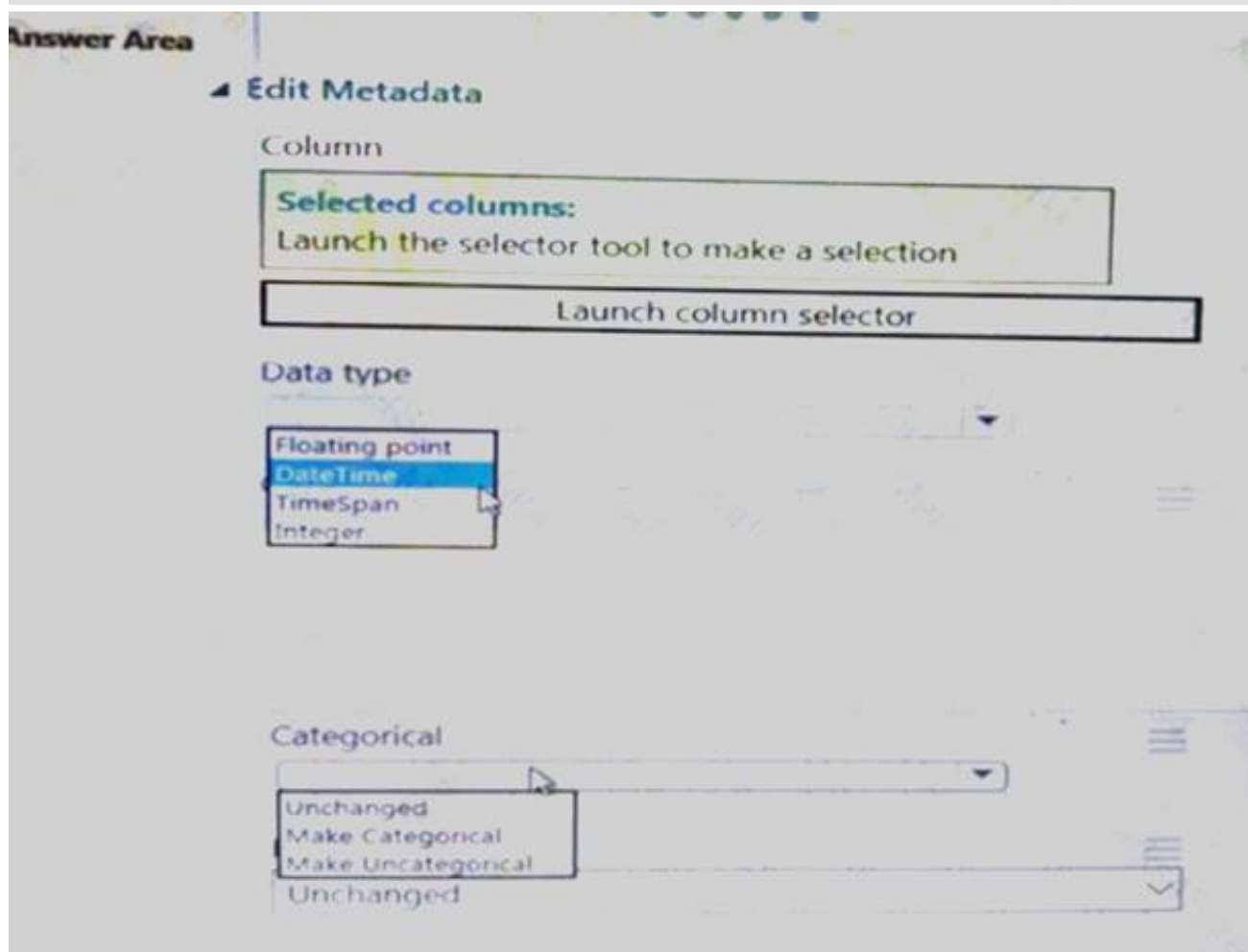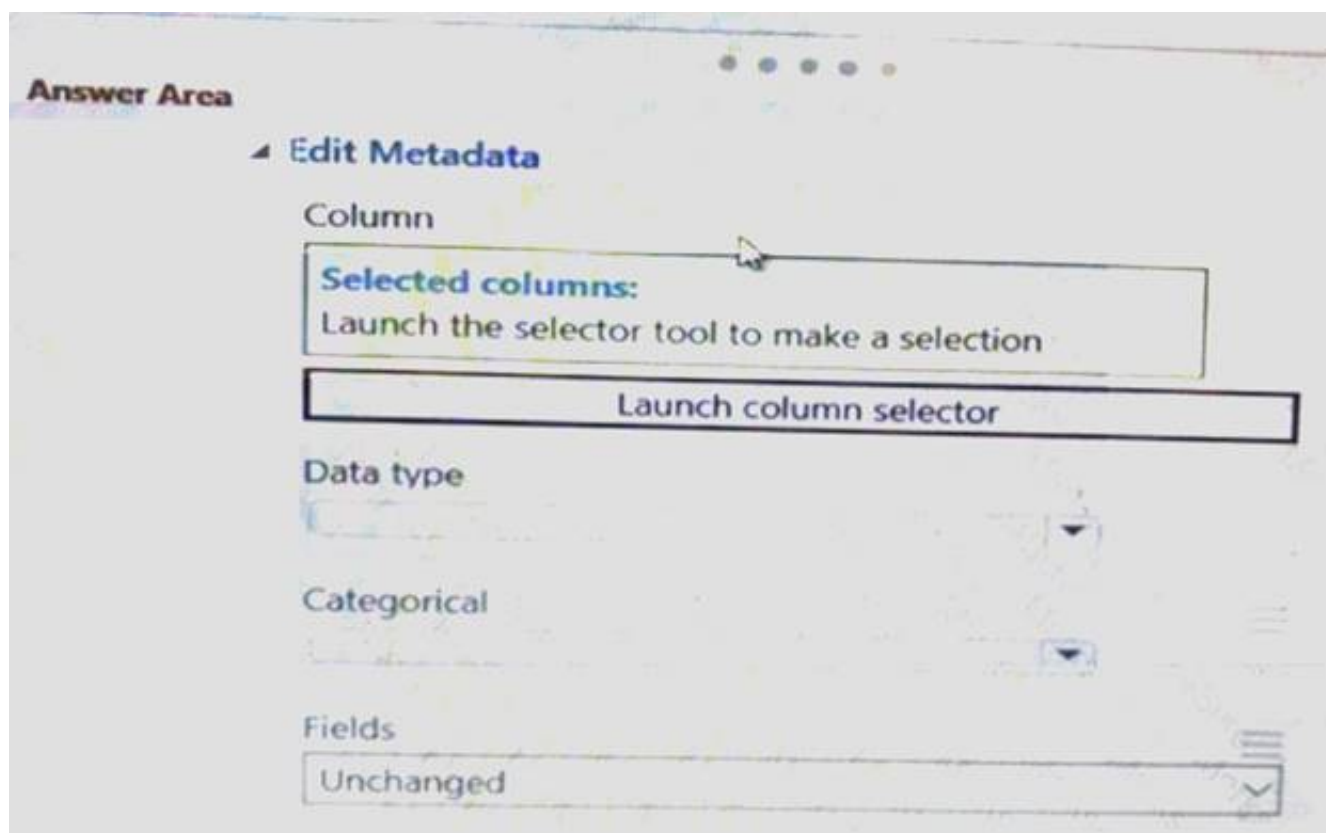

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Step 1: Augment the data
Scenario: Columns in each dataset contain missing and null values. The datasets also contain many outliers.
Step 2: Add the Bayesian Linear Regression module.
Scenario: You produce a regression model to predict property prices by using the Linear Regression and Bayesian Linear Regression modules.
Step 3: Configure the regularization weight.
Regularization typically is used to avoid overfitting. For example, in L2 regularization weight, type the value to use as the weight for L2 regularization. We recommend that you use a non-zero value to avoid overfitting.
Scenario:
Model fit: The model shows signs of overfitting. You need to produce a more refined regression model that reduces the overfitting.
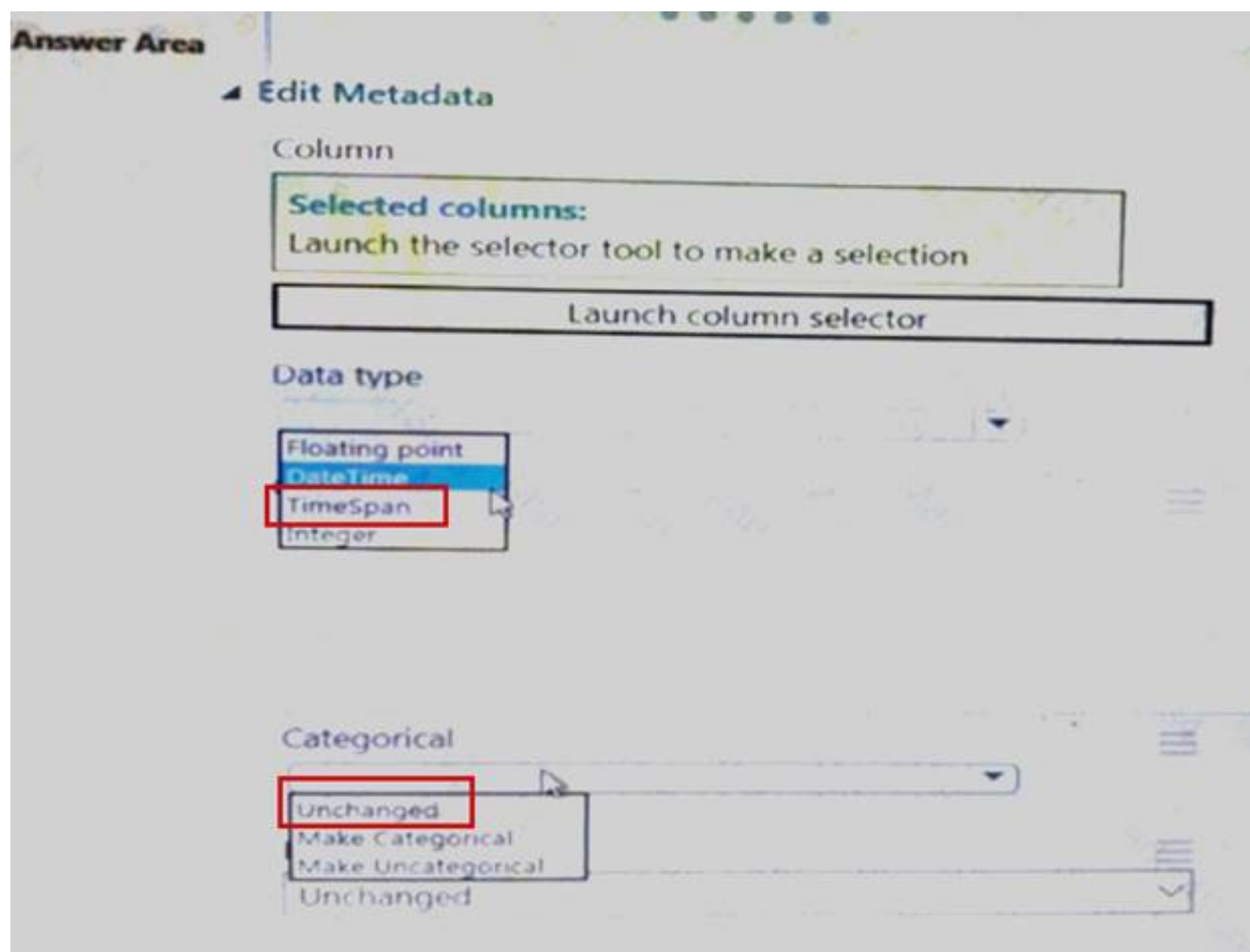
NEW QUESTION 23
- (Exam Topic 2)
You need to configure the Edit Metadata module so that the structure of the datasets match. Which configuration options should you select? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.

Answer Area

▲ Edit Metadata

Column

Selected columns:
Launch the selector tool to make a selection

Launch column selector

Data type

▼

Categorical

▼

Fields
Unchanged

Answer Area

▲ Edit Metadata

Column

Selected columns:
Launch the selector tool to make a selection

Launch column selector

Data type

Floating point
DateTime
TimeSpan
Integer

Categorical

▼

Unchanged
Make Categorical
Make Uncategorical
Unchanged

A. Mastered
B. Not Mastered

Answer: A

Explanation:

Answer Area

◢ **Edit Metadata**

Column

Selected columns:
Launch the selector tool to make a selection

Launch column selector

Data type

| Floating point |
| DateTime |
| TimeSpan |
| Integer |

Categorical

| Unchanged |
| Make Categorical |
| Make Uncategorical |

Unchanged

NEW QUESTION 24
- (Exam Topic 2)
You need to configure the Feature Based Feature Selection module based on the experiment requirements and datasets.
How should you configure the module properties? To answer, select the appropriate options in the dialog box in the answer area.
NOTE: Each correct selection is worth one point.

◢ Filter Based Feature Selection

Feature scoring method

| Fisher Score |
| Chi-squared |
| Mutual information |
| Counts |

☑ Operate on feature columns only

Target column

| MedianValue |
| AvgRooms/nHouse |

Launch column selector

Number of desired features

1

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Box 1: Mutual Information.
The mutual information score is particularly useful in feature selection because it maximizes the mutual information between the joint distribution and target variables in datasets with many dimensions.
Box 2: MedianValue
MedianValue is the feature column, , it is the predictor of the dataset.
Scenario: The MedianValue and AvgRoomsinHouse columns both hold data in numeric format. You need to select a feature selection algorithm to analyze the relationship between the two columns in more detail.

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

NEW QUESTION 27
- (Exam Topic 2)
You need to implement early stopping criteria as suited in the model training requirements.

Which three code segments should you use to develop the solution? To answer, move the appropriate code segments from the list of code segments to the answer area and arrange them in the correct order.

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

Code segments	Answer Area

```
early_termination_policy =
TruncationSelectionPolicy(evaluation_interval=1,
truncation_percentage=20, delay_evaluation=5)
```

```
import TruncationSelectionPolicy
```

```
from azureml.train.hyperdrive
```

```
import BanditPolicy
```

```
early_termination_policy = BanditPolicy
(slack_factor = 0.1, evaluation_interval=1,
delay_evaluation=5)
```

A. Mastered
B. Not Mastered

Answer: A

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

NEW QUESTION 31
- (Exam Topic 2)
You need to replace the missing data in the AccessibilityToHighway columns.
How should you configure the Clean Missing Data module? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.

Properties   Project

◢ Clean Missing Data

Columns to be cleaned

Selected columns:
Column names: AccessibilityToHighway

Launch column selector

Minimum missing value ratio

0

Maximum missing value ratio

1

Cleaning mode

▼

Replace using MICE
Replace with Mean
Replace with Median
Replace with Mode

Cols with all missing values.

▼

Propagate
Remove

☑ Generate missing value indicator column

Number of iterations

5

A. Mastered
B. Not Mastered

Answer: A

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

NEW QUESTION 33
- (Exam Topic 1)
You need to define an evaluation strategy for the crowd sentiment models.
Which three actions should you perform in sequence? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.

Actions

Add new features for retraining supervised models.

Filter labeled cases for retraining using the shortest distance from centroids.

Evaluate the changes in correlation between model error rate and centroid distance

Impute unavailable features with centroid aligned models

Filter labeled cases for retraining using the longest distance from centroids.

Remove features before retraining supervised models.

Answer Area

A. Mastered
B. Not Mastered

Answer: A

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

NEW QUESTION 37
- (Exam Topic 1)
You need to define a process for penalty event detection.
Which three actions should you perform in sequence? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.

Actions

Build the global model using Microsoft Cognitive Toolkit (CNTK).

Import the global model and build a local model using Microsoft Cognitive Toolkit (CNTK).

Export the global model using Neural Network Exchange Format (NNEF).

Import the global model and build the local model using PyTorch.

Build the global model using PyTorch.

Build the global model using TensorFlow.

Import the global model and build the local model using TensorFlow.

Export the global model using the Open Neural Network Exchange (ONNX) format.
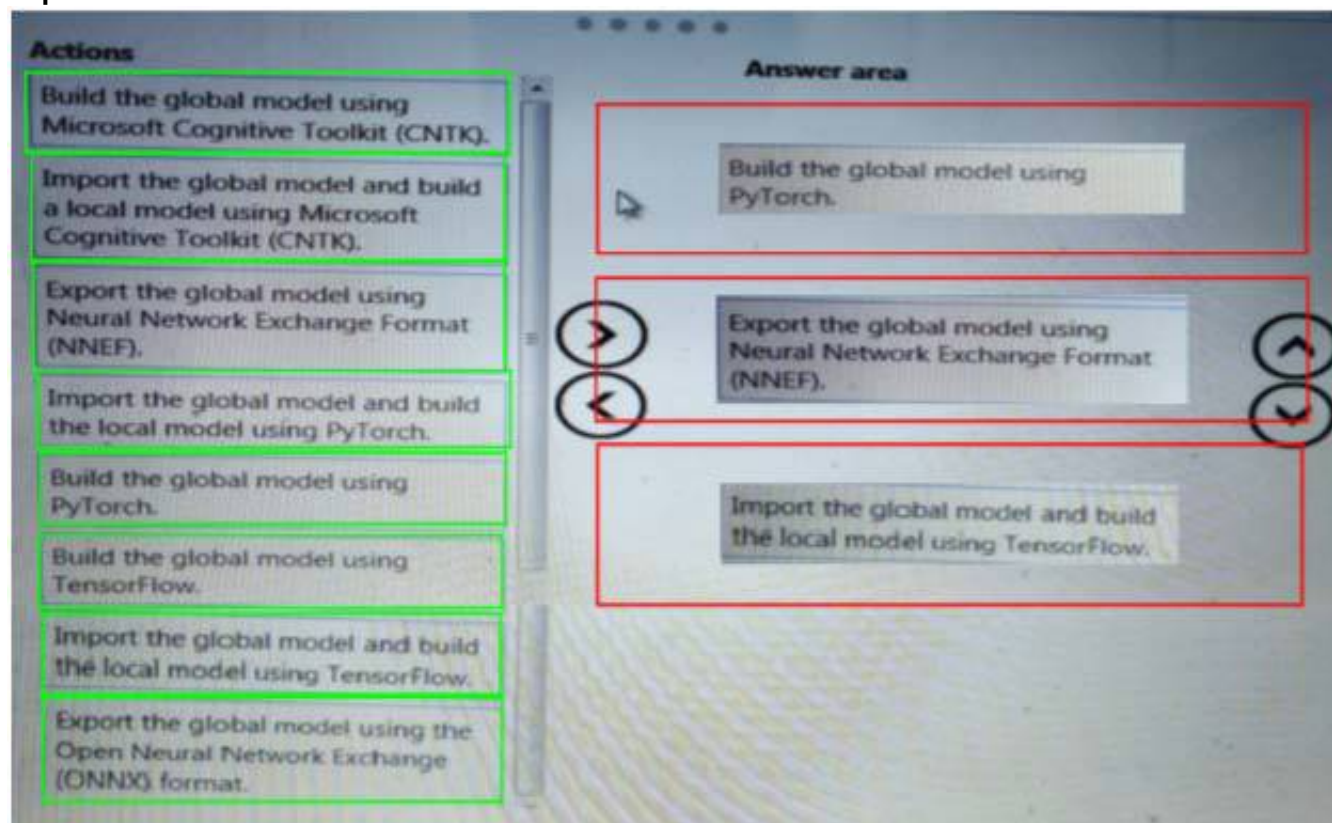
Answer area

A. Mastered
B. Not Mastered
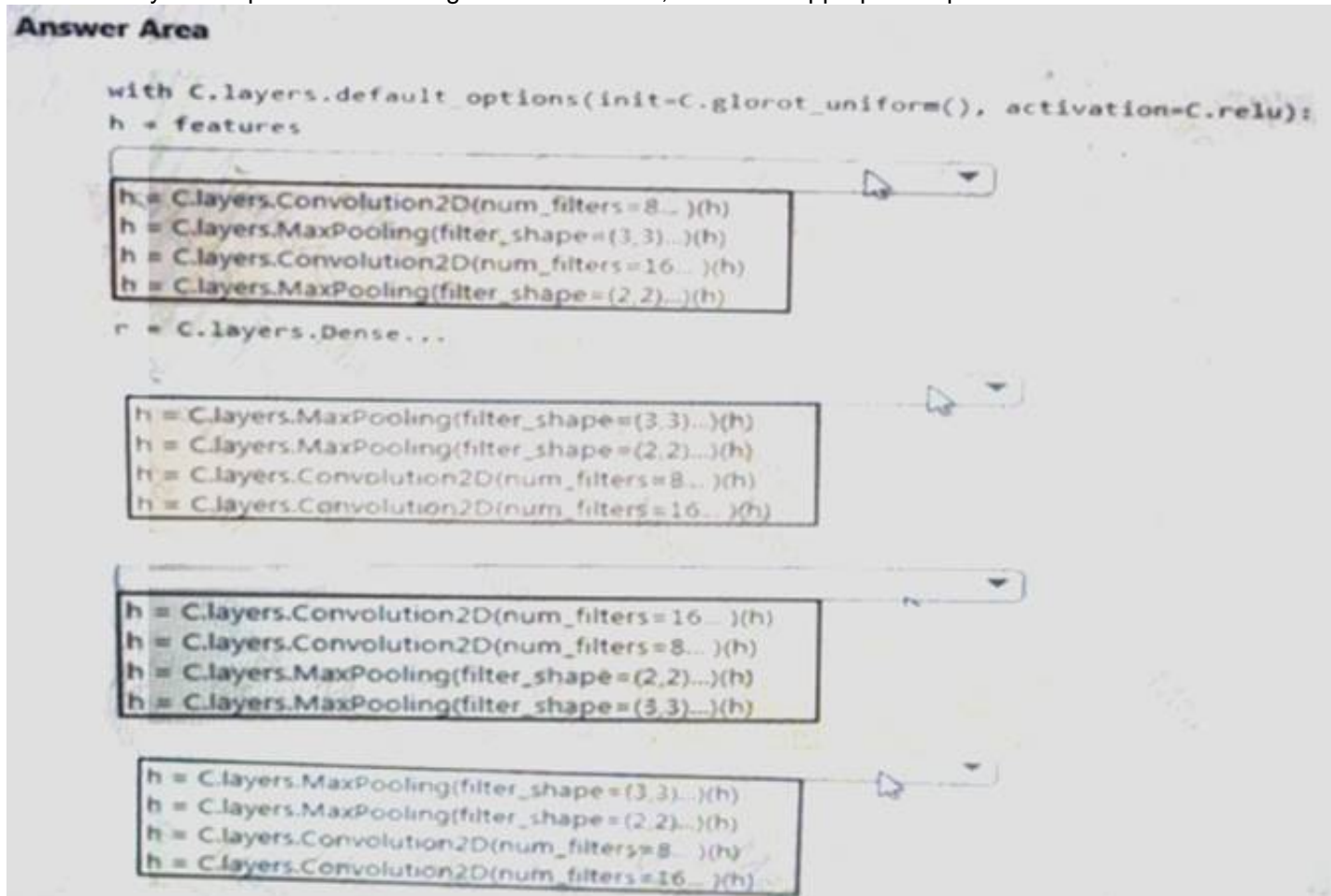
Answer: A

Explanation:

NEW QUESTION 41
- (Exam Topic 1)
You need to build a feature extraction strategy for the local models.
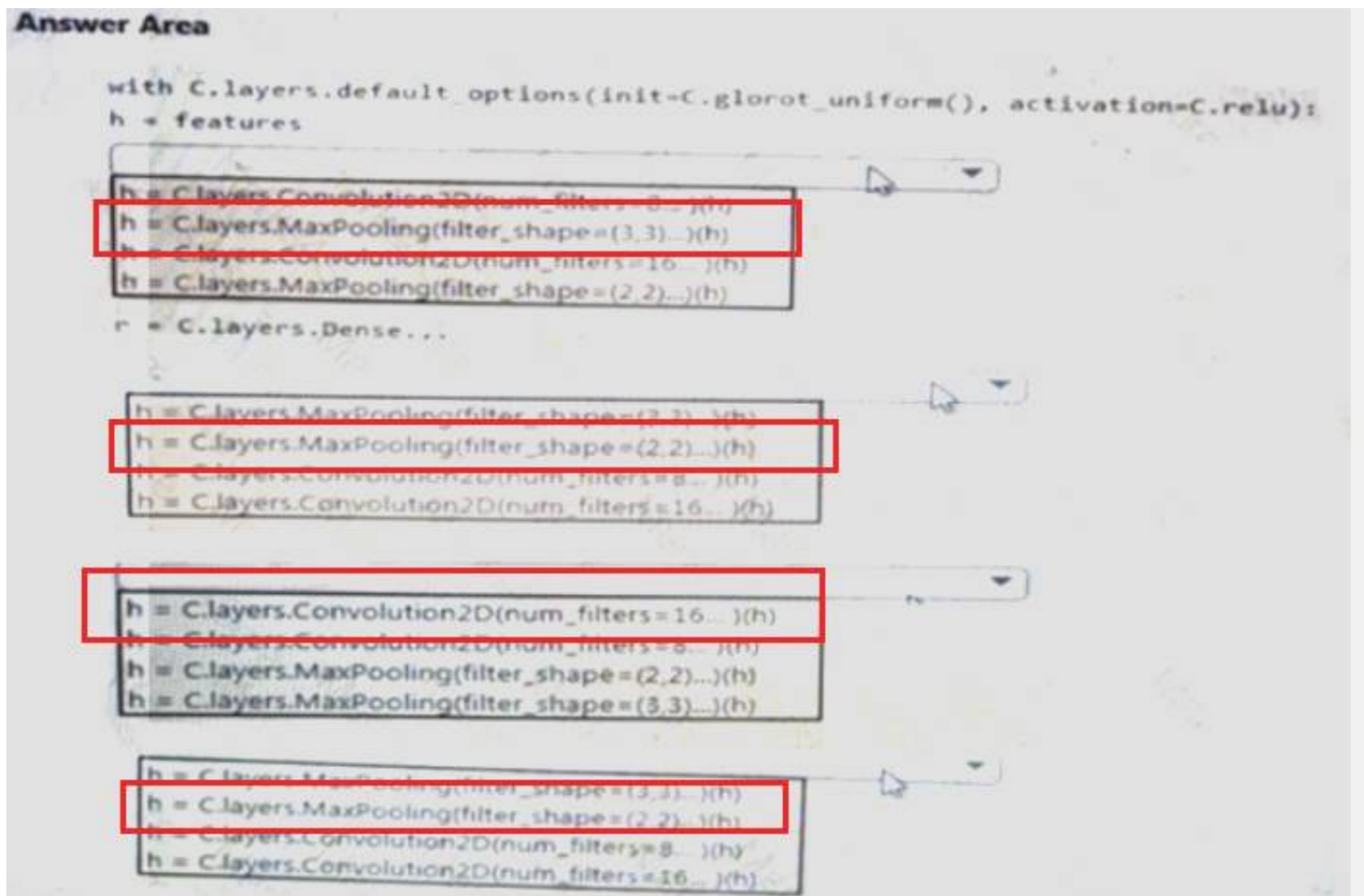How should you complete the code segment? To answer, select the appropriate options in the answer area. NOTE: Each correct selection is worth one point.

A. Mastered
B. Not Mastered

Answer: A

Explanation:

Answer Area

```
with C.layers.default_options(init=C.glorot_uniform(), activation=C.relu):
    h = features
```
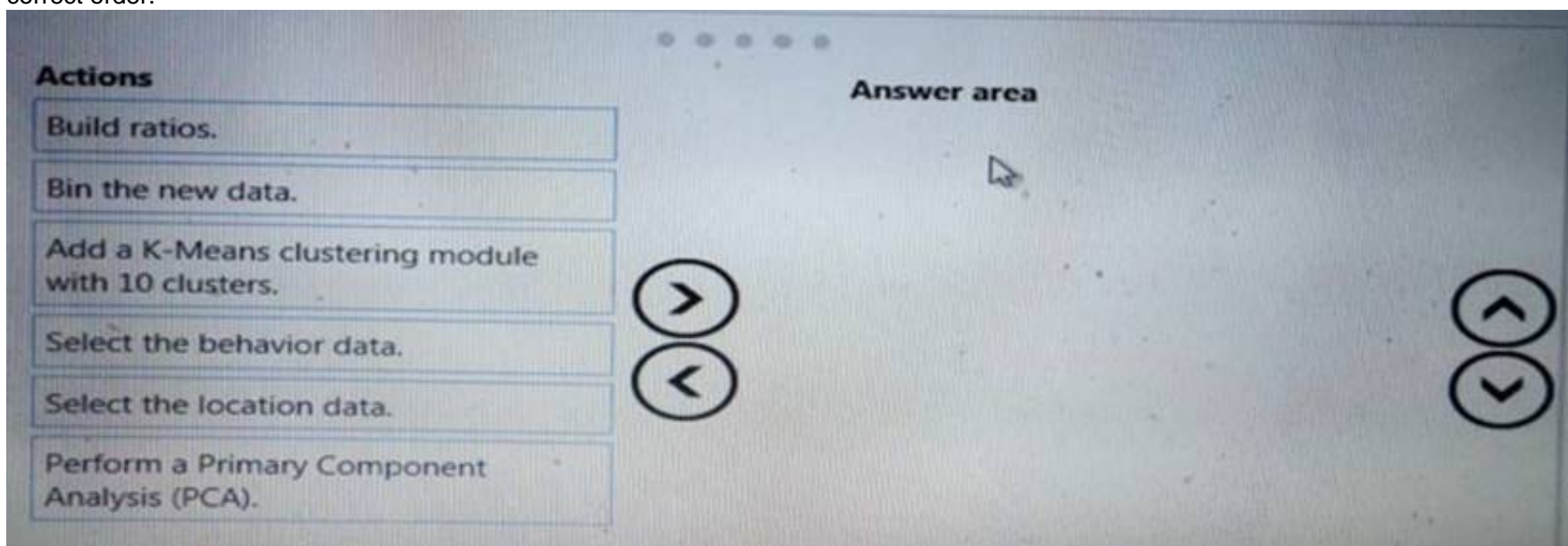


```
    r = C.layers.Dense...
```


NEW QUESTION 46
- (Exam Topic 1)
You need to modify the inputs for the global penalty event model to address the bias and variance issue.
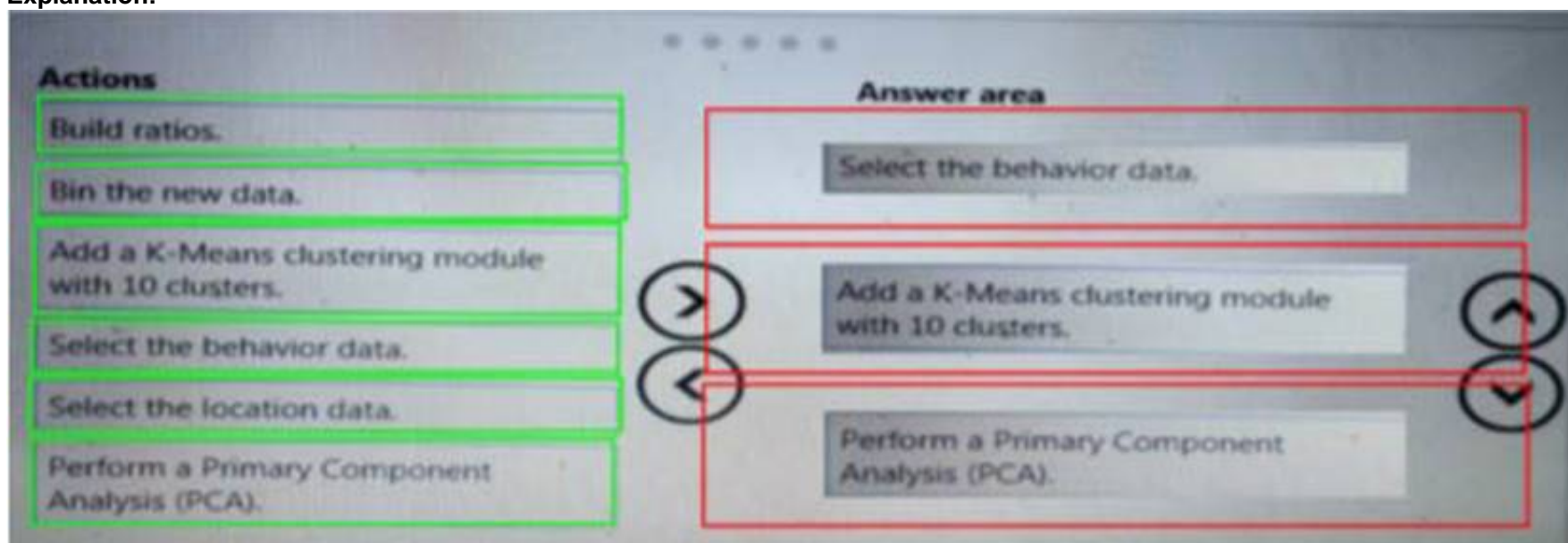Which three actions should you perform in sequence? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.

A. Mastered
B. Not Mastered

Answer: A

Explanation:

NEW QUESTION 49

- (Exam Topic 3)
You are developing a hands-on workshop to introduce Docker for Windows to attendees. You need to ensure that workshop attendees can install Docker on their devices.
Which two prerequisite components should attendees install on the devices? Each correct answer presents part of the solution.
NOTE: Each correct selection is worth one point.

A. Microsoft Hardware-Assisted Virtualization Detection Tool
B. Kitematic
C. BIOS-enabled virtualization
D. VirtualBox
E. Windows 10 64-bit Professional

Answer: E

Explanation:
C: Make sure your Windows system supports Hardware Virtualization Technology and that virtualization is enabled.
Ensure that hardware virtualization support is turned on in the BIOS settings. For example:

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

NEW QUESTION 51
- (Exam Topic 3)
You are performing sentiment analysis using a CSV file that includes 12,000 customer reviews written in a short sentence format. You add the CSV file to Azure Machine Learning Studio and configure it as the starting point dataset of an experiment. You add the Extract N-Gram Features from Text module to the experiment to extract key phrases from the customer review column in the dataset.
You must create a new n-gram dictionary from the customer review text and set the maximum n-gram size to trigrams.
What should you select? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.

Properties   Project

Extract N-Gram Features from Text

Text column

Selected columns:
Column type: String Feature

Launch column selector

Vocabulary mode

Create
ReadOnly
Update
Merge

N-Grams size

3
4
4,000
12,000

0

Weighting function

Minimum word length

3

Maximum word length

25

Minimum n-gram document absolu...

5

Maximum n-gram document ratio

1

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Vocabulary mode: Create
For Vocabulary mode, select Create to indicate that you are creating a new list of n-gram features. N-Grams size: 3
For N-Grams size, type a number that indicates the maximum size of the n-grams to extract and store. For example, if you type 3, unigrams, bigrams, and trigrams will be created.
Weighting function: Leave blank
The option, Weighting function, is required only if you merge or update vocabularies. It specifies how terms in the two vocabularies and their scores should be weighted against each other.
References:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/extract-n-gram-features-from

NEW QUESTION 54
- (Exam Topic 3)
You create a classification model with a dataset that contains 100 samples with Class A and 10,000 samples with Class B
The variation of Class B is very high. You need to resolve imbalances. Which method should you use?

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

Answer: D

NEW QUESTION 58
- (Exam Topic 3)
You plan to build a team data science environment. Data for training models in machine learning pipelines will be over 20 GB in size.
You have the following requirements:

≫ Models must be built using Caffe2 or Chainer frameworks.

≫ Data scientists must be able to use a data science environment to build the machine learning pipelines and train models on their personal devices in both connected and disconnected network environments.

≫ Personal devices must support updating machine learning pipelines when connected to a network. You need to select a data science environment.
Which environment should you use?

A. Azure Machine Learning Service
B. Azure Machine Learning Studio
C. Azure Databricks
D. Azure Kubernetes Service (AKS)

Answer: A

Explanation:
The Data Science Virtual Machine (DSVM) is a customized VM image on Microsoft's Azure cloud built specifically for doing data science. Caffe2 and Chainer are supported by DSVM.
DSVM integrates with Azure Machine Learning.

NEW QUESTION 62
- (Exam Topic 3)
You are performing a classification task in Azure Machine Learning Studio.
You must prepare balanced testing and training samples based on a provided data set. You need to split the data with a 0.75:0.25 ratio.
Which value should you use for each parameter? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.

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

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Box 1: Split rows
Use the Split Rows option if you just want to divide the data into two parts. You can specify the percentage of data to put in each split, but by default, the data is divided 50-50.
You can also randomize the selection of rows in each group, and use stratified sampling. In stratified sampling, you must select a single column of data for which you want values to be apportioned equally among the two result datasets.
Box 2: 0.75
If you specify a number as a percentage, or if you use a string that contains the "%" character, the value is interpreted as a percentage. All percentage values must be within the range (0, 100), not including the values 0 and 100.
Box 3: Yes
To ensure splits are balanced. Box 4: No
If you use the option for a stratified split, the output datasets can be further divided by subgroups, by selecting a strata column.
Reference:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/split-data

NEW QUESTION 63
- (Exam Topic 3)
You are developing a data science workspace that uses an Azure Machine Learning service. You need to select a compute target to deploy the workspace.
What should you use?

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

Answer: D

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

NEW QUESTION 64
- (Exam Topic 3)
Your team is building a data engineering and data science development environment. The environment must support the following requirements:

> support Python and Scala

> compose data storage, movement, and processing services into automated data pipelines

> the same tool should be used for the orchestration of both data engineering and data science

> support workload isolation and interactive workloads

> enable scaling across a cluster of machines You need to create the environment.
What should you do?

A. Build the environment in Apache Hive for HDInsight and use Azure Data Factory for orchestration.
B. Build the environment in Azure Databricks and use Azure Data Factory for orchestration.
C. Build the environment in Apache Spark for HDInsight and use Azure Container Instances for orchestration.
D. Build the environment in Azure Databricks and use Azure Container Instances for orchestration.

Answer: B

Explanation:
In Azure Databricks, we can create two different types of clusters.

> Standard, these are the default clusters and can be used with Python, R, Scala and SQL

> High-concurrency
Azure Databricks is fully integrated with Azure Data Factory.

NEW QUESTION 67
- (Exam Topic 3)
You are a data scientist creating a linear regression model.
You need to determine how closely the data fits the regression line. Which metric should you review?

A. Coefficient of determination
B. Recall
C. Precision
D. Mean absolute error
E. Root Mean Square Error

Answer: A

Explanation:
Coefficient of determination, often referred to as R2, represents the predictive power of the model as a value between 0 and 1. Zero means the model is random (explains nothing); 1 means there is a perfect fit. However, caution should be used in interpreting R2 values, as low values can be entirely normal and high values can be suspect.
References:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/evaluate-model

NEW QUESTION 70
- (Exam Topic 3)
You are developing deep learning models to analyze semi-structured, unstructured, and structured data types. You have the following data available for model building:

> Video recordings of sporting events

> Transcripts of radio commentary about events

> Logs from related social media feeds captured during sporting events You need to select an environment for creating the model.
Which environment should you use?

A. Azure Cognitive Services
B. Azure Data Lake Analytics
C. Azure HDInsight with Spark MLib
D. Azure Machine Learning Studio

Answer: A

Explanation:
Azure Cognitive Services expand on Microsoft's evolving portfolio of machine learning APIs and enable developers to easily add cognitive features – such as emotion and video detection; facial, speech, and vision recognition; and speech and language understanding – into their applications. The goal of Azure Cognitive Services is to help developers create applications that can see, hear, speak, understand, and even begin to reason. The catalog of services within Azure Cognitive

Services can be categorized into five main pillars - Vision, Speech, Language, Search, and Knowledge.
References:
https://docs.microsoft.com/en-us/azure/cognitive-services/welcome

NEW QUESTION 71
- (Exam Topic 3)
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.
You are analyzing a numerical dataset which contains missing values in several columns.
You must clean the missing values using an appropriate operation without affecting the dimensionality of the feature set.
You need to analyze a full dataset to include all values.
Solution: Replace each missing value using the Multiple Imputation by Chained Equations (MICE) method. Does the solution meet the goal?

A. Yes
B. NO

Answer: A

Explanation:
Replace using MICE: For each missing value, this option assigns a new value, which is calculated by using a method described in the statistical literature as "Multivariate Imputation using Chained Equations" or "Multiple Imputation by Chained Equations". With a multiple imputation method, each variable with missing data is modeled conditionally using the other variables in the data before filling in the missing values.
Note: Multivariate imputation by chained equations (MICE), sometimes called "fully conditional specification" or "sequential regression multiple imputation" has emerged in the statistical literature as one principled method of addressing missing data. Creating multiple imputations, as opposed to single imputations, accounts for the statistical uncertainty in the imputations. In addition, the chained equations approach is very flexible and can handle variables of varying types (e.g., continuous or binary) as well as complexities such as bounds or survey skip patterns.
References: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3074241/
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/clean-missing-data

NEW QUESTION 76
- (Exam Topic 3)
You are developing a linear regression model in Azure Machine Learning Studio. You run an experiment to compare different algorithms.
The following image displays the results dataset output:

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

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

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

A. Mastered
B. Not Mastered

Answer: A

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

NEW QUESTION 80
- (Exam Topic 3)
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.

◢ SMOTE

Label column

Selected columns:
All labels

Launch column selector

SMOTE percentage

▼
0
300
3000
4000

Number of nearest neighbors

▼
0
1
5
4000

Random seed

0

A. Mastered
B. Not Mastered

Answer: A

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

NEW QUESTION 84
- (Exam Topic 3)
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.
You are using Azure Machine Learning Studio to perform feature engineering on a dataset. You need to normalize values to produce a feature column grouped into bins.
Solution: Apply an Entropy Minimum Description Length (MDL) binning mode.
Does the solution meet the goal?

A. Yes

B. No

Answer: A

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

NEW QUESTION 87
- (Exam Topic 3)
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.
You are a data scientist using Azure Machine Learning Studio.
You need to normalize values to produce an output column into bins to predict a target column. Solution: Apply a Quantiles binning mode with a PQuantile normalization.
Does the solution meet the goal?

A. Yes
B. No

Answer: B

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

NEW QUESTION 90
- (Exam Topic 3)
You plan to preprocess text from CSV files. You load the Azure Machine Learning Studio default stop words list.
You need to configure the Preprocess Text module to meet the following requirements:

> Ensure that multiple related words from a single canonical form.

> Remove pipe characters from text.

> Remove words to optimize information retrieval.

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

◢ Preprocess Text

Language

| English | ∨ |

Remove by part of speech

| False | ∨ |

Text column to clean

Selected columns:
Column names: **String, Feature**

| Launch column selector |

☐ Remove stop words	≡
☐ Lemmatization	≡
☐ Detect sentences	≡
☐ Normalize case to lowercase	≡
☐ Remove numbers	≡
☐ Remove special characters	≡
☐ Remove duplicate characters	≡
☐ Remove email addresses	≡
☐ Remove URLs	≡
☐ Expand verb contractions	≡
☐ Normalize backslashes to slashes	≡
☐ Split tokens on special characters	≡

A. Mastered
B. Not Mastered

Answer: A

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

NEW QUESTION 94
- (Exam Topic 3)
You need to select a feature extraction method. Which method should you use?

A. Spearman correlation
B. Mutual information
C. Mann-Whitney test
D. Pearson's correlation

Answer: D

NEW QUESTION 96
- (Exam Topic 3)
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.
You are analyzing a numerical dataset which contains missing values in several columns.
You must clean the missing values using an appropriate operation without affecting the dimensionality of the feature set.
You need to analyze a full dataset to include all values.
Solution: Remove the entire column that contains the missing data point. Does the solution meet the goal?

A. Yes
B. No

Answer: B

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

NEW QUESTION 98
- (Exam Topic 3)
You are performing clustering by using the K-means algorithm. You need to define the possible termination conditions.
Which three conditions can you use? Each correct answer presents a complete solution. NOTE: Each correct selection is worth one point.

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

Answer: ADE

Explanation:
 References:
https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/k-means-clustering https://nlp.stanford.edu/IR-book/html/htmledition/k-means-1.html

NEW QUESTION 102
- (Exam Topic 3)
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.
You are creating a new experiment in Azure Machine Learning Studio.
One class has a much smaller number of observations than the other classes in the training set. You need to select an appropriate data sampling strategy to compensate for the class imbalance. Solution: You use the Scale and Reduce sampling mode.
Does the solution meet the goal?

A. Yes
B. No

Answer: B

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

NEW QUESTION 105
- (Exam Topic 3)
You create a binary classification model by using Azure Machine Learning Studio.
You must tune hyperparameters by performing a parameter sweep of the model. The parameter sweep must meet the following requirements:
> iterate all possible combinations of hyperparameters
> minimize computing resources required to perform the sweep
> You need to perform a parameter sweep of the model.
Which parameter sweep mode should you use?

A. Random sweep
B. Sweep clustering
C. Entire grid
D. Random grid
E. Random seed

Answer: D

Explanation:
Maximum number of runs on random grid: This option also controls the number of iterations over a random sampling of parameter values, but the values are not

generated randomly from the specified range; instead, a matrix is created of all possible combinations of parameter values and a random sampling is taken over the matrix. This method is more efficient and less prone to regional oversampling or undersampling.

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

NEW QUESTION 106
- (Exam Topic 3)
You are building a machine learning model for translating English language textual content into French language textual content.
You need to build and train the machine learning model to learn the sequence of the textual content. Which type of neural network should you use?

A. Multilayer Perceptions (MLPs)
B. Convolutional Neural Networks (CNNs)
C. Recurrent Neural Networks (RNNs)
D. Generative Adversarial Networks (GANs)

Answer: C

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

NEW QUESTION 111
- (Exam Topic 3)
You are creating a new experiment in Azure Machine Learning Studio. You have a small dataset that has missing values in many columns. The data does not require the application of predictors for each column. You plan to use the Clean Missing Data module to handle the missing data.
You need to select a data cleaning method. Which method should you use?

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

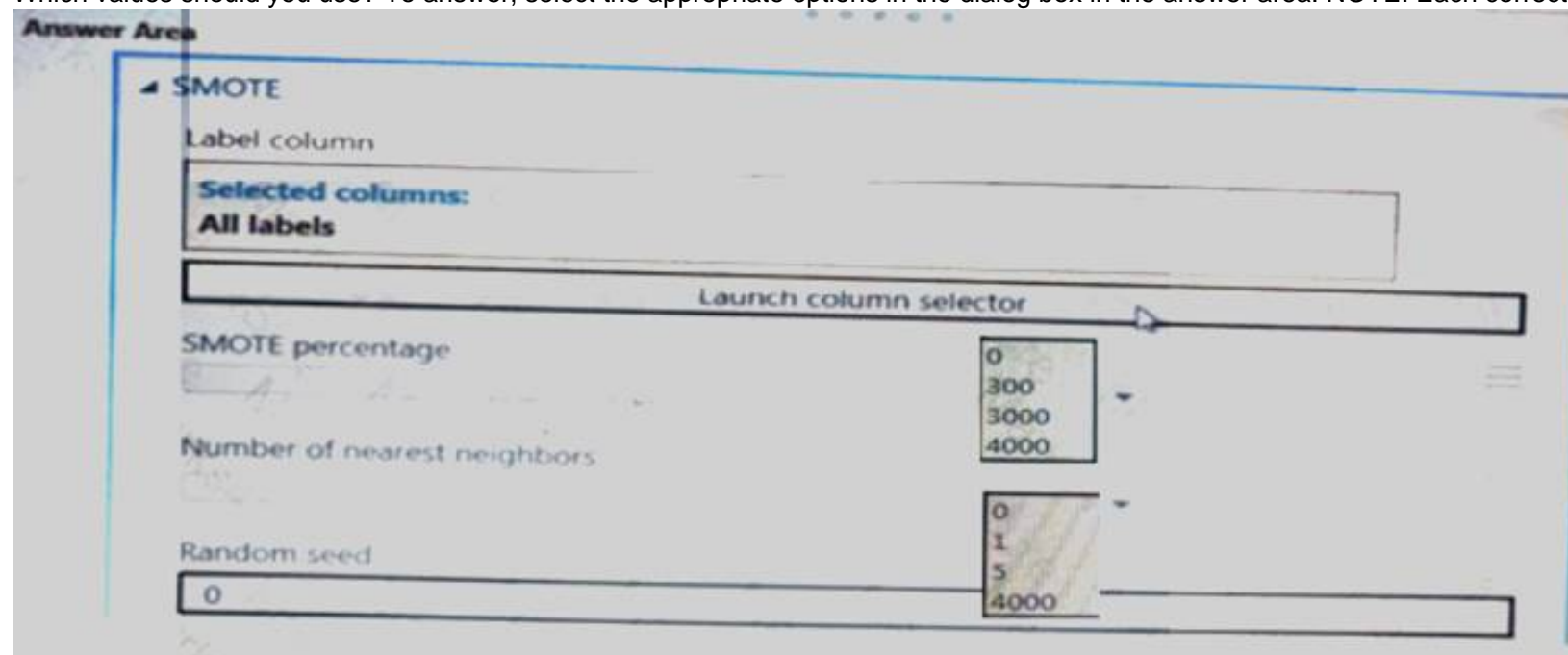**Answer:** A

NEW QUESTION 115
- (Exam Topic 3)
You create an experiment in Azure Machine Learning Studio- You add a training dataset that contains 10.000 rows. The first 9.000 rows represent class 0 (90 percent). The first 1.000 rows represent class 1 (10 percent).
The training set is unbalanced between two Classes. You must increase the number of training examples for class 1 to 4,000 by using data rows. You add the Synthetic Minority Oversampling Technique (SMOTE) module to the experiment.
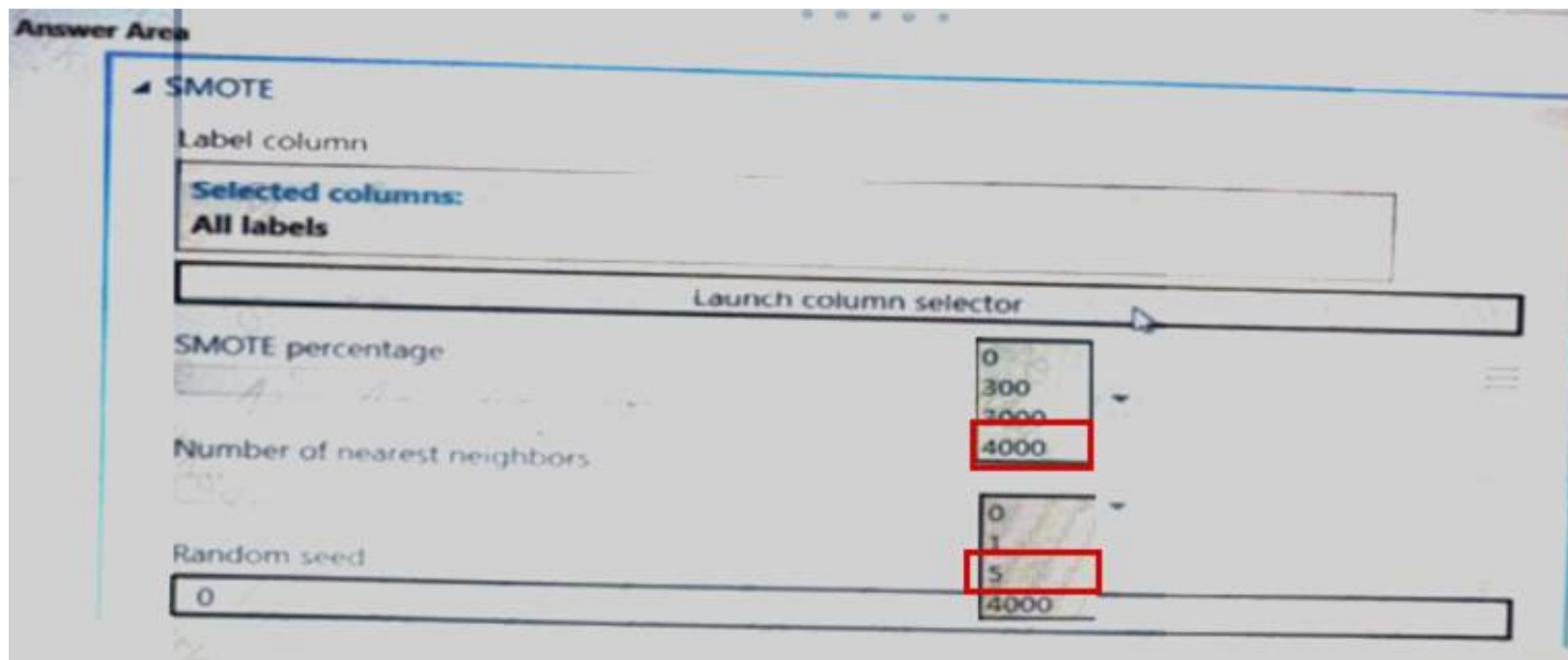You need to configure the module.
Which values should you use? To answer, select the appropriate options in the dialog box in the answer area. NOTE: Each correct selection is worth one point.

A. Mastered
B. Not Mastered

Answer: A

Explanation:

NEW QUESTION 120
- (Exam Topic 3)
You are creating a binary classification by using a two-class logistic regression model. You need to evaluate the model results for imbalance.
Which evaluation metric should you use?

A. Relative Absolute Error
B. AUC Curve
C. Mean Absolute Error
D. Relative Squared Error

Answer: B

Explanation:
One can inspect the true positive rate vs. the false positive rate in the Receiver Operating Characteristic (ROC) curve and the corresponding Area Under the Curve (AUC) value. The closer this curve is to the upper left corner, the better the classifier's performance is (that is maximizing the true positive rate while minimizing the false positive rate). Curves that are close to the diagonal of the plot, result from classifiers that tend to make predictions that are close to random guessing.
References:
https://docs.microsoft.com/en-us/azure/machine-learning/studio/evaluate-model-performance#evaluating-a-bina

NEW QUESTION 125
- (Exam Topic 3)
You are creating a machine learning model in Python. The provided dataset contains several numerical columns and one text column. The text column represents a product's category. The product category will always be one of the following:
➢ Bikes
➢ Cars
➢ Vans
➢ Boats
You are building a regression model using the scikit-learn Python package.
You need to transform the text data to be compatible with the scikit-learn Python package.
How should you complete the code segment? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.

```
from sklearn import linear_model
import  [▼]
        pandas as df
        numpy as df
        scipy as df

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

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

A. Mastered
B. Not Mastered

Answer: A

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

NEW QUESTION 128
- (Exam Topic 3)
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.

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

A. Mastered
B. Not Mastered

Answer: A

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

NEW QUESTION 131
- (Exam Topic 3)
You are implementing a machine learning model to predict stock prices. The model uses a PostgreSQL database and requires GPU processing.
You need to create a virtual machine that is pre-configured with the required tools. What should you do?

A. Create a Data Science Virtual Machine (DSVM) Windows edition.
B. Create a Geo AI Data Science Virtual Machine (Geo-DSVM) Windows edition.
C. Create a Deep Learning Virtual Machine (DLVM) Linux edition.
D. Create a Deep Learning Virtual Machine (DLVM) Windows edition.
E. Create a Data Science Virtual Machine (DSVM) Linux edition.

Answer: E

NEW QUESTION 136
......

Thank You for Trying Our Product

* 100% Pass or Money Back

    All our products come with a 90-day Money Back Guarantee.

* One year free update

    You can enjoy free update one year. 24x7 online support.

* Trusted by Millions

    We currently serve more than 30,000,000 customers.

* Shop Securely

    All transactions are protected by VeriSign!

100% Pass Your DP-100 Exam with Our Prep Materials Via below:

https://www.certleader.com/DP-100-dumps.html