

Microsoft

Exam Questions DP-100

Designing and Implementing a Data Science Solution on Azure



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.

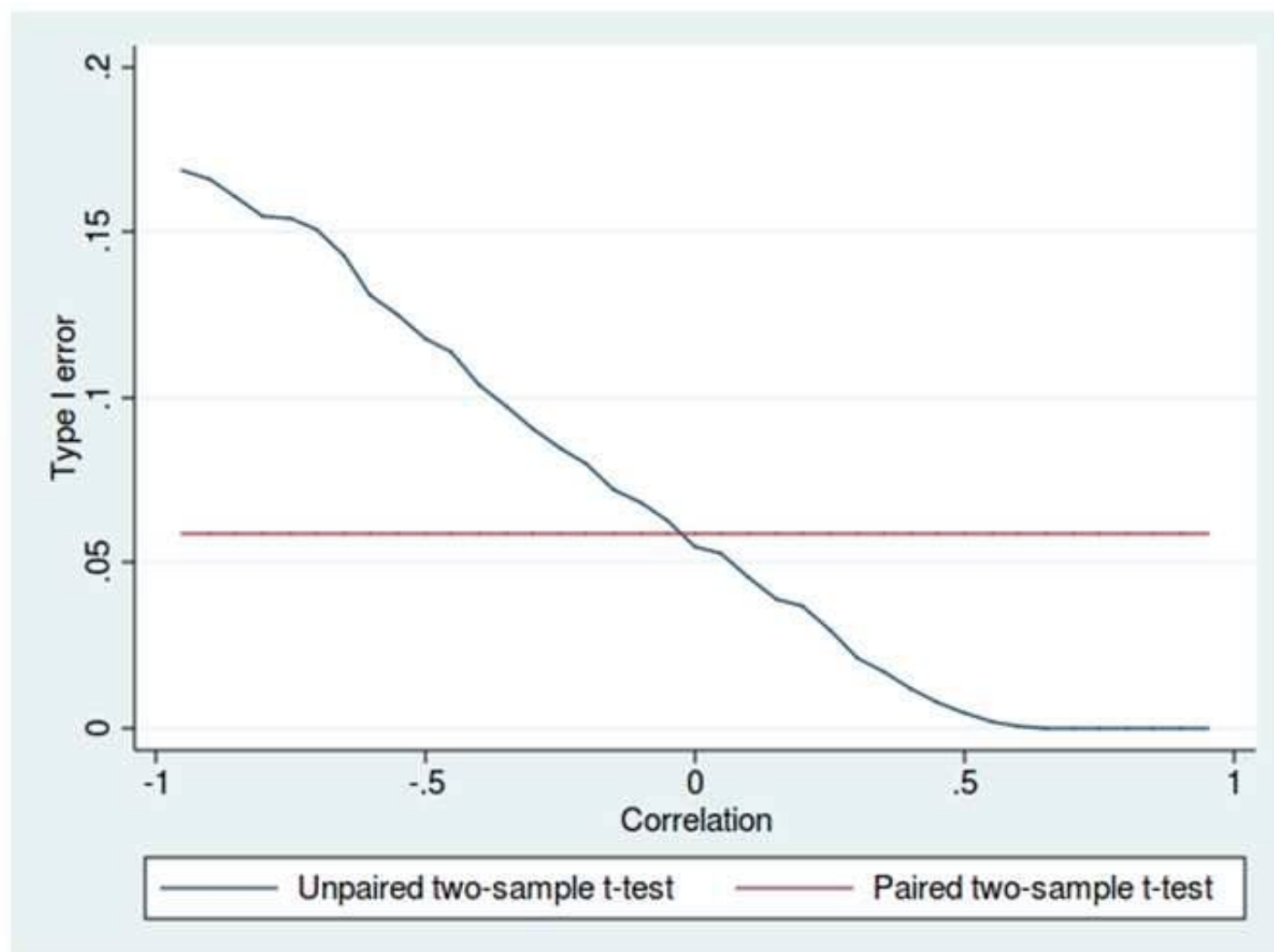
Which type of distribution should you produce?

- 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)

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: $\text{ZeroOneLoss}(x,y) = 1$ when $x \neq y$; otherwise 0.

Coefficient of determination, often referred to as R^2 , 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 3

- (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 4

- (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 5

- (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 6

- (Exam Topic 3)

You are creating a machine learning model. You have a dataset that contains null rows.

You need to use the Clean Missing Data module in Azure Machine Learning Studio to identify and resolve the null and missing data in the dataset. Which parameter should you use?

- A. Replace with mean
- B. Remove entire column
- C. Remove entire row
- D. Hot Deck

Answer: B

Explanation:

Remove entire row: Completely removes any row in the dataset that has one or more missing values. This is useful if the missing value can be considered randomly missing.

References:

<https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/clean-missing-data>

NEW QUESTION 7

- (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 τ), 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 8

- (Exam Topic 2)

You need to produce a visualization for the diagnostic test evaluation according to the data visualization requirements.

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

Modules		Answer Area
Score Matchbox Recommender		
Apply Transformation		
Evaluate Recommender		
Evaluate Model	⬅	⬆
Train Model	➡	⬇
Sweep Clustering		
Score Model		
Load Trained Model		

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Step 1: Sweep Clustering

Start by using the "Tune Model Hyperparameters" module to select the best sets of parameters for each of the models we're considering.

One of the interesting things about the "Tune Model Hyperparameters" module is that it not only outputs the results from the Tuning, it also outputs the Trained Model.

Step 2: Train Model Step 3: Evaluate Model

Scenario: You need to provide the test results to the Fabrikam Residences team. You create data visualizations to aid in presenting the results.

You must produce a Receiver Operating Characteristic (ROC) curve to conduct a diagnostic test evaluation of the model. You need to select appropriate methods for producing the ROC curve in Azure Machine Learning Studio to compare the Two-Class Decision Forest and the Two-Class Decision Jungle modules with one another.

References:

<http://breaking-bi.blogspot.com/2017/01/azure-machine-learning-model-evaluation.html>

NEW QUESTION 9

- (Exam Topic 2)

You need to configure the Permutation Feature Importance module for the model training requirements. What should you do? To answer, select the appropriate options in the dialog box in the answer area. NOTE: Each correct selection is worth one point.

Answer Area

Permutation Feature importance

Random seed

▼

0

500

▼

Regression – Root Mean Square Error

Regression – R-squared

Regression – Mean Zero One Error

Regression – Mean Absolute Error

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Box 1: 500
For Random seed, type a value to use as seed for randomization. If you specify 0 (the default), a number is generated based on the system clock. A seed value is optional, but you should provide a value if you want reproducibility across runs of the same experiment. Here we must replicate the findings. Box 2: Mean Absolute Error
Scenario: Given a trained model and a test dataset, you must compute the Permutation Feature Importance scores of feature variables. You need to set up the Permutation Feature Importance module to select the correct metric to investigate the model's accuracy and replicate the findings. Regression. Choose one of the following: Precision, Recall, Mean Absolute Error , Root Mean Squared Error, Relative Absolute Error, Relative Squared Error, Coefficient of Determination
References:
<https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/permutation-feature-importan>

NEW QUESTION 10

- (Exam Topic 2)
You need to identify the methods for dividing the data according, to the testing requirements. Which properties should you select? To answer, select the appropriate option-, m the answer area. NOTE: Each correct selection is worth one point.

Answer Area

Properties Project

Partition and Sample

Partition or sample mode

Assign to Folds
Sampling
Head

Random seed

0

Specify the partitioner method

Partition evenly

Specify number of folds to split evenly into

5

Stratified split

Classification key column

Selected columns:

Column names: NextToRiver

Launch column selector

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Answer Area

Properties Project

Partition and Sample

Partition or sample mode

Assign to Folds
Sampling
Head

Random seed

0

Specify the partitioner method

Partition evenly

Specify number of folds to split evenly into

5

Stratified split

Classification key column

Selected columns:

Column names: NextToRiver

Launch column selector

NEW QUESTION 10
- (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

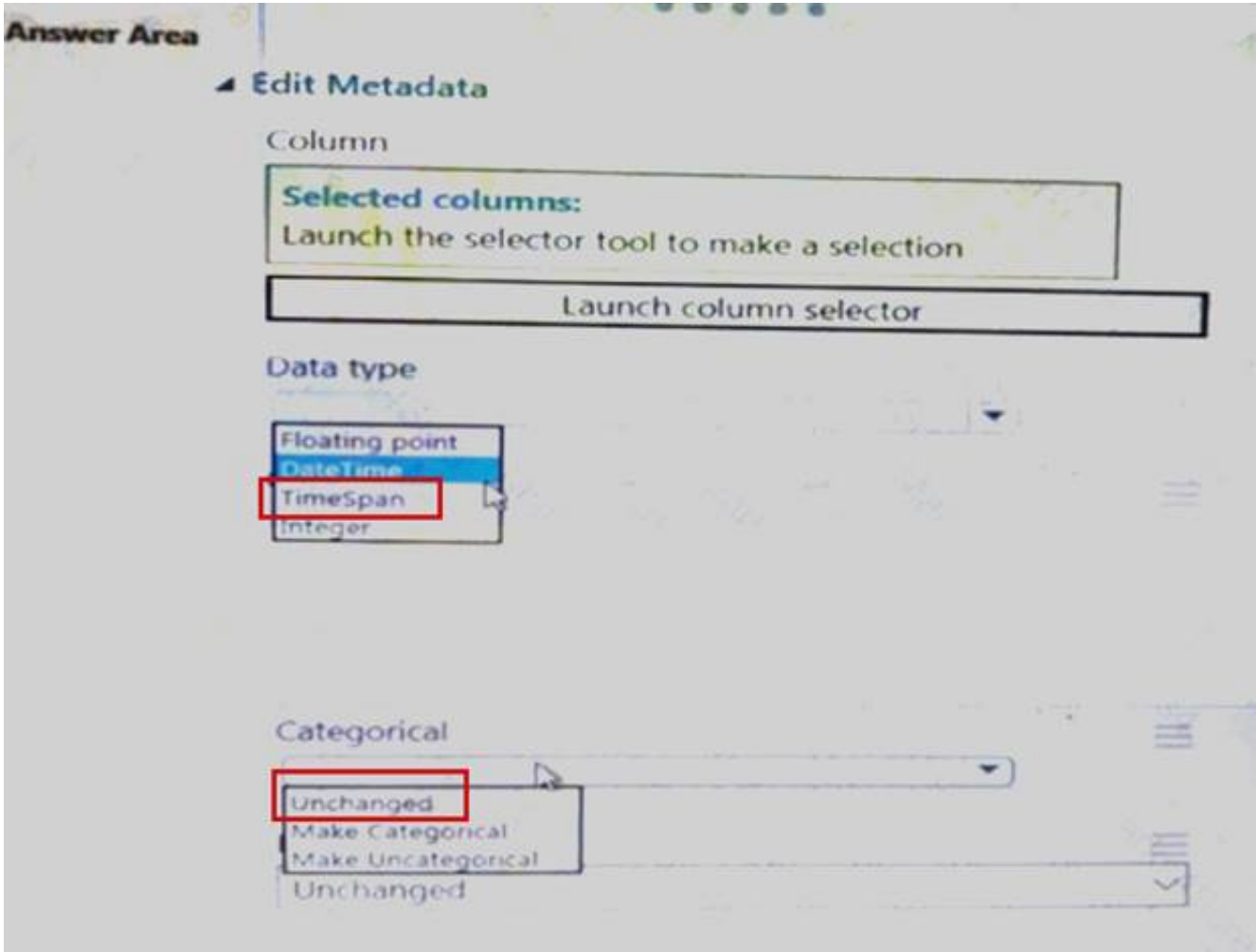
Categorical

Unchanged
 Make Categorical
 Make Uncategorical
 Unchanged

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:



NEW QUESTION 14

- (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		Answer Area
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.		

- 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

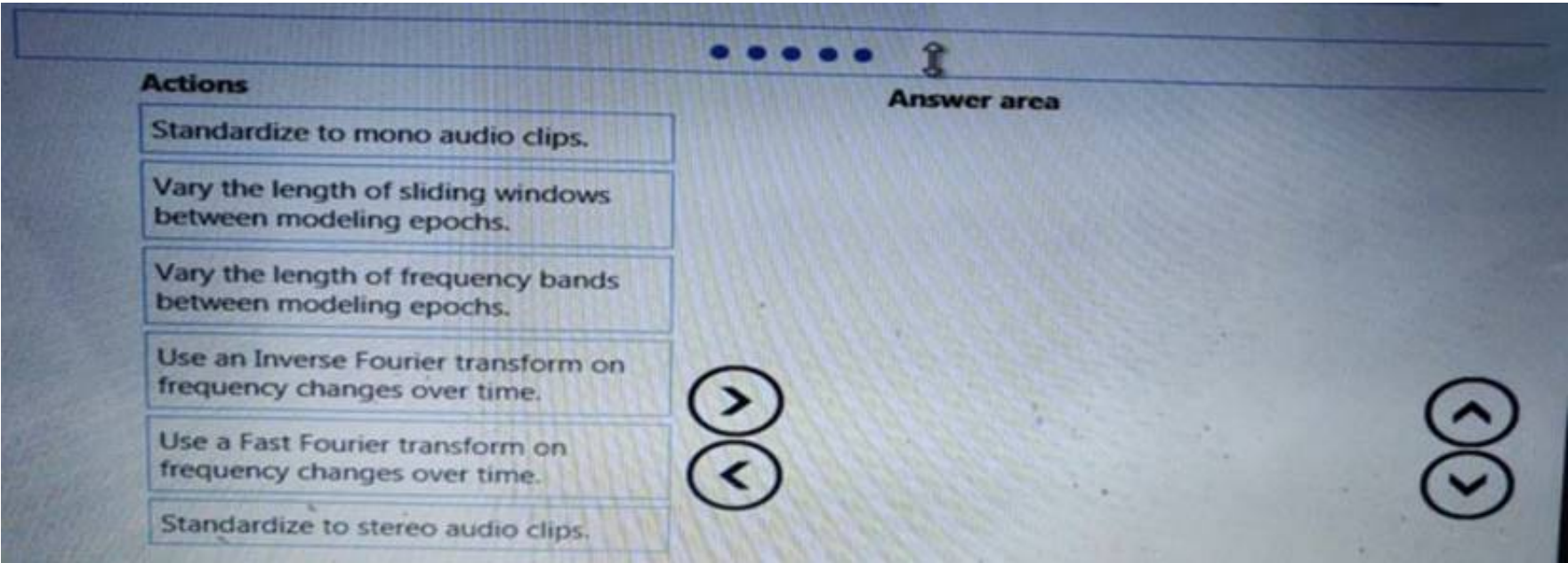
<https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/sweep-clustering>

NEW QUESTION 16

- (Exam Topic 1)

You need to define a process for penalty event detection.

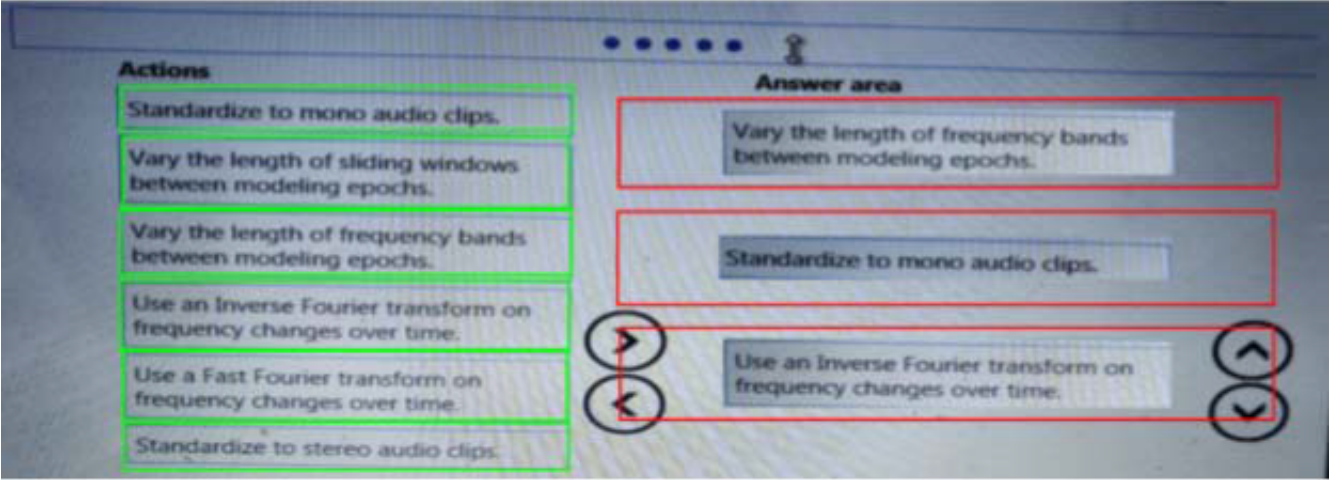
Which three actions should you perform in sequence? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.



- A. Mastered
- B. Not Mastered

Answer: A

Explanation:



NEW QUESTION 20

- (Exam Topic 3)
You have a Python data frame named salesData in the following format: The data frame must be unpivoted to a long data format as follows:
You need to use the pandas.melt() function in Python to perform the transformation.
How should you complete the code segment? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.

Answer Area

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

dataFrame

pandas

salesData

year

, id_vars='

shop

year

value

Shop X, Shop Y, Shop Z

, value_vars=

'shop'

'year'

['year']

['2017', '2018']

)

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

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

2 c B 5
3 a C 2
4 b C 4
5 c C 6

References:

<https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.melt.html>

NEW QUESTION 24

- (Exam Topic 3)

You must store data in Azure Blob Storage to support Azure Machine Learning. You need to transfer the data into Azure Blob Storage.

What are three possible ways to achieve the goal? Each correct answer presents a complete solution.

NOTE: Each correct selection is worth one point.

- A. Bulk Insert SQL Query
- B. AzCopy
- C. Python script
- D. Azure Storage Explorer
- E. Bulk Copy Program (BCP)

Answer: BCD

Explanation:

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

AzCopy Python SSIS

References:

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

NEW QUESTION 29

- (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 31

- (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	<div>▼ Split rows Recommender Split Regular Expression Split Relative Expression Split</div>
Fraction of rows in the first output dataset	<div>▼ 0.75 0.25 0.5 1</div>
Randomized split	<div>▼ True False</div>
Stratified split	<div>▼ True False</div>

- 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 33

- (Exam Topic 3)
You are using C-Support Vector classification to do a multi-class classification with an unbalanced training dataset. The C-Support Vector classification using Python code shown below:

```
from sklearn.svm import svc
import numpy as np
svc = SVC(kernel= 'linear', class_weight= 'balanced', C=1.0, random_state=0)
modell = svc.fit(X_train, y)
```

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

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

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Box 1: Automatically adjust weights inversely proportional to class frequencies in the input data

The “balanced” mode uses the values of y to automatically adjust weights inversely proportional to class frequencies in the input data as $n_{\text{samples}} / (n_{\text{classes}} * \text{np.bincount}(y))$.

Box 2: Penalty parameter

Parameter: C : float, optional (default=1.0)

Penalty parameter C of the error term. References:

<https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html>

NEW QUESTION 35

- (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: Relative Squared Error, Coefficient of Determination, Accuracy, Precision, Recall, F1 score, and AUC.

Does the solution meet the goal?

- A. Yes
- B. No

Answer: B

Explanation:

Relative Squared Error, Coefficient of Determination are good metrics to evaluate the linear regression model, but the others are metrics for classification models.

References:

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

NEW QUESTION 39

- (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: Accuracy, Precision, Recall, F1 score and AUC. Does the solution meet the goal?

- A. Yes
- B. No

Answer: B

Explanation:

Those are metrics for evaluating classification models, instead use: Mean Absolute Error, Root Mean Absolute Error, Relative Absolute Error, Relative Squared Error, and the Coefficient of Determination.

References:

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

NEW QUESTION 41

- (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 46

- (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 MLlib
- 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 50

- (Exam Topic 3) You are solving a classification task. The dataset is imbalanced.

You need to select an Azure Machine Learning Studio module to improve the classification accuracy. Which module should you use?

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

Answer: C

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>

NEW QUESTION 53

- (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 55

- (Exam Topic 3)

You are evaluating a Python NumPy array that contains six data points defined as follows: data = [10, 20, 30, 40, 50, 60]

You must generate the following output by using the k-fold algorithm implantation in the Python Scikit-learn machine learning library:

train: [10 40 50 60], test: [20 30]

train: [20 30 40 60], test: [10 50]

train: [10 20 30 50], test: [40 60]

You need to implement a cross-validation to generate the output.

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

NOTE: Each correct selection is worth one point.

```
from numpy import array
from sklearn.model_selection import K-Means
k-fold
CrossValidation
ModelSelection

data = array([10, 20, 30, 40, 50, 60])
kfold = Kfold(n_splits=1
2
3
6, shuffle = True, random_state=1)

for train, test in kFold, split(data
k-fold
array
train, test):

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

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Box 1: k-fold

Box 2: 3

K-F olds cross-validator provides train/test indices to split data in train/test sets. Split dataset into k consecutive folds (without shuffling by default).

The parameter n_splits (int, default=3) is the number of folds. Must be at least 2. Box 3: data

Example: Example:

>>>

>>> from sklearn.model_selection import KFold

>>> X = np.array([[1, 2], [3, 4], [1, 2], [3, 4]])

>>> y = np.array([1, 2, 3, 4])

>>> kf = KFold(n_splits=2)

>>> kf.get_n_splits(X) 2

>>> print(kf)

KFold(n_splits=2, random_state=None, shuffle=False)

>>> for train_index, test_index in kf.split(X): print("TRAIN:", train_index, "TEST:", test_index) X_train, X_test = X[train_index], X[test_index] y_train, y_test =

y[train_index], y[test_index] TRAIN: [2 3] TEST: [0 1]

TRAIN: [0 1] TEST: [2 3]

References:

https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.KFold.html

NEW QUESTION 60

- (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 61

- (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 65

- (Exam Topic 3)

You have a dataset created for multiclass classification tasks that contains a normalized numerical feature set with 10,000 data points and 150 features.

You use 75 percent of the data points for training and 25 percent for testing. You are using the scikit-learn machine learning library in Python. You use X to denote the feature set and Y to denote class labels.

You create the following Python data frames:

You need to apply the Principal Component Analysis (PCA) method to reduce the dimensionality of the feature set to 10 features in both training and testing sets. How should you complete the code segment? To answer, select the appropriate options in the answer area. NOTE: Each correct selection is worth one point.

```
from sklearn.decomposition import PCA
pca = 
    PCA()
    PCA(n_components = 150)
    PCA(n_components = 10)
    PCA(n_components = 10000)

X_train = 
    pca
    model
    sklearn.decomposition
.fit_transform(X_train)
x_test = pca.
    x_test
    X_train
    fit(x_test)
    transform(x_test)

```

- A. Mastered
- B. Not Mastered

Answer: A

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. References:

<https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html>

NEW QUESTION 70

- (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 71

- (Exam Topic 3)

You are evaluating a completed binary classification machine learning model. You need to use the precision as the valuation metric.

Which visualization should you use?

- A. Binary classification confusion matrix
- B. box plot
- C. Gradient descent
- D. coefficient of determination

Answer: A

Explanation:

References:

<https://machinelearningknowledge.ai/confusion-matrix-and-performance-metrics-machine-learning/>

NEW QUESTION 76

.....

Thank You for Trying Our Product

We offer two products:

1st - We have Practice Tests Software with Actual Exam Questions

2nd - Questions and Answers in PDF Format

DP-100 Practice Exam Features:

- * DP-100 Questions and Answers Updated Frequently
- * DP-100 Practice Questions Verified by Expert Senior Certified Staff
- * DP-100 Most Realistic Questions that Guarantee you a Pass on Your FirstTry
- * DP-100 Practice Test Questions in Multiple Choice Formats and Updatesfor 1 Year

100% Actual & Verified — Instant Download, Please Click
[Order The DP-100 Practice Test Here](#)