

# 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](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  $R^2$  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 tuning a hyperparameter for an algorithm. The following table shows a data set with different hyperparameter, training error, and validation errors.

Use the drop-down menus to select the answer choice that answers each question based on the information presented in the graphic.

- 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 4**

- (Exam Topic 3)

You are creating an experiment by using Azure Machine Learning Studio.

You must divide the data into four subsets for evaluation. There is a high degree of missing values in the data. You must prepare the data for analysis.

You need to select appropriate methods for producing the experiment.

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

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

- A. Mastered
- B. Not Mastered

**Answer:** A

**Explanation:**

The Clean Missing Data module in Azure Machine Learning Studio, to remove, replace, or infer missing values.

**NEW QUESTION 5**

- (Exam Topic 3)

You are analyzing a dataset containing historical data from a local taxi company. You are developing a regression model.

You must predict the fare of a taxi trip.

You need to select performance metrics to correctly evaluate the regression model. Which two metrics can you use? Each correct answer presents a complete solution. NOTE: Each correct selection is worth one point.

- A. an F1 score that is high
- B. an R Squared value close 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 low
- F. an F1 score that is low.

**Answer:** DF

**NEW QUESTION 6**

- (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.

- 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 7**

- (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](https://en.wikipedia.org/wiki/Convolutional_neural_network)

**NEW QUESTION 8**

- (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 normalization with a QuantileIndex 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 9**

- (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  $\tau$ ), 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 10**

- (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 10**

- (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.

- 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 14**

- (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.

- 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 19**

- (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.

- 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 21**

- (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.

- 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 24**

- (Exam Topic 1)

You need to resolve the local machine learning pipeline performance issue. What should you do?

- A. Increase Graphic Processing Units (GPUs).
- B. Increase the learning rate.
- C. Increase the training iterations,
- D. Increase Central Processing Units (CPUs).

**Answer:** A

**NEW QUESTION 29**

- (Exam Topic 1)

You need to implement a scaling strategy for the local penalty detection data. Which normalization type should you use?

- A. Streaming
- B. Weight
- C. Batch
- D. Cosine

**Answer:** C

**Explanation:**

Post batch normalization statistics (PBN) is the Microsoft Cognitive Toolkit (CNTK) version of how to evaluate the population mean and variance of Batch Normalization which could be used in inference Original Paper.

In CNTK, custom networks are defined using the BrainScriptNetworkBuilder and described in the CNTK network description language "BrainScript."

Scenario:

Local penalty detection models must be written by using BrainScript. References:

<https://docs.microsoft.com/en-us/cognitive-toolkit/post-batch-normalization-statistics>

**NEW QUESTION 31**

- (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 36**

- (Exam Topic 1)

You need to select an environment that will meet the business and data requirements. Which environment should you use?

- A. Azure HDInsight with Spark MLlib
- B. Azure Cognitive Services
- C. Azure Machine Learning Studio
- D. Microsoft Machine Learning Server

**Answer:** D

**NEW QUESTION 41**

- (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 43**

- (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.

- 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

Parameter 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 47**

- (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 Learning learning Studio.

One class has a much smaller number of observations than the other classes in the training

You need to select an appropriate data sampling strategy to compensate for the class imbalance. Solution: You use the Synthetic Minority Oversampling Technique (SMOTE) sampling mode. Does the solution meet the goal?

- A. Yes
- B. No

**Answer:** A

**Explanation:**

SMOTE is used 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.

References:

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

**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://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 52**

- (Exam Topic 3)

You plan to use a Data Science Virtual Machine (DSVM) with the open source deep learning frameworks Caffe2 and Theano. You need to select a pre configured DSVM to support the framework.

What should you create?

- A. Data Science Virtual Machine for Linux (CentOS)
- B. Data Science Virtual Machine for Windows 2012
- C. Data Science Virtual Machine for Windows 2016

- D. Geo AI Data Science Virtual Machine with ArcGIS
- E. Data Science Virtual Machine for Linux (Ubuntu)

**Answer:** E

**NEW QUESTION 57**

- (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.

- 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 58**

- (Exam Topic 3)

You are analyzing the asymmetry in a statistical distribution.

The following image contains two density curves that show the probability distribution of two datasets.

Use the drop-down menus to select the answer choice that answers each question based on the information presented in the graphic.

NOTE: Each correct selection is worth one point.

- A. Mastered
- B. Not Mastered

**Answer:** A

**Explanation:**

Box 1: Positive skew

Positive skew values means the distribution is skewed to the right. Box 2: Negative skew

Negative skewness values mean the distribution is skewed to the left. References:

<https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/compute-elementary-statistic>

**NEW QUESTION 61**

- (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 64**

- (Exam Topic 3)

You create a binary classification model using Azure Machine Learning Studio.

You must use a Receiver Operating Characteristic (ROC) curve and an F1 score to evaluate the model. You need to create the required business metrics.

How should you complete the experiment? 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 65**

- (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.

- 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 70**

- (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:

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.

- 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 72**

- (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 77**

- (Exam Topic 3)

You are performing a filter based feature selection for a dataset 10 build a multi class classifiers by using Azure Machine Learning Studio.

The dataset contains categorical features that are highly correlated to the output label column.

You need to select the appropriate feature scoring statistical method to identify the key predictors. Which method should you use?

- A. Chi-squared

- B. Spearman correlation
- C. Kendall correlation
- D. Person correlation

**Answer:** D

**Explanation:**

Pearson's correlation statistic, or Pearson's correlation coefficient, is also known in statistical models as the r value. For any two variables, it returns a value that indicates the strength of the correlation  
 Pearson's correlation coefficient is the test statistics that measures the statistical relationship, or association, between two continuous variables. It is known as the best method of measuring the association between variables of interest because it is based on the method of covariance. It gives information about the magnitude of the association, or correlation, as well as the direction of the relationship.

Reference:

<https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/filter-based-feature-selection> <https://www.statisticssolutions.com/pearsons-correlation-coefficient/>

**NEW QUESTION 78**

- (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 80**

- (Exam Topic 3)

You need to select a pre built development environment for a series of data science experiments. You must use the R language for the experiments.

Which three environments can you use? Each correct answer presents a complete solution. NOTE: Each correct selection is worth one point.

- A. MI.NET Library on a local environment
- B. Azure Machine Learning Studio
- C. Data Science Virtual Machine (OSVM)
- D. Azure Data bricks
- E. Azure Cognitive Services

**Answer:** ABD

**NEW QUESTION 84**

- (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 88**

- (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:

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.

- 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 91**

- (Exam Topic 3)

You are using a decision tree algorithm. You have trained a model that generalizes well at a tree depth equal to 10.

You need to select the bias and variance properties of the model with varying tree depth values.

Which properties should you select for each tree depth? To answer, select the appropriate options in the answer area.

- A. Mastered
- B. Not Mastered

**Answer:** A

**Explanation:**

In decision trees, the depth of the tree determines the variance. A complicated decision tree (e.g. deep) has low bias and high variance.

Note: In statistics and machine learning, the bias–variance tradeoff is the property of a set of predictive models whereby models with a lower bias in parameter estimation have a higher variance of the parameter estimates across samples, and vice versa. Increasing the bias will decrease the variance. Increasing the variance will decrease the bias.

References:

<https://machinelearningmastery.com/gentle-introduction-to-the-bias-variance-trade-off-in-machine-learning/>

**NEW QUESTION 92**

- (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 94**

- (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.

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

**NEW QUESTION 99**

- (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.

- 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 104**

- (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 107**

- (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 111**

- (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 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.

References:

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

**NEW QUESTION 113**

- (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 118**

- (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.

- 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 122**

- (Exam Topic 3)

You plan to use a Deep Learning Virtual Machine (DLVM) to train deep learning models using Compute Unified Device Architecture (CUDA) computations.

You need to configure the DLVM to support CUDA. What should you implement?

- A. Intel Software Guard Extensions (Intel SGX) technology
- B. Solid State Drives (SSD)
- C. Graphic Processing Unit (GPU)
- D. Computer Processing Unit (CPU) speed increase by using overclocking
- E. High Random Access Memory (RAM) configuration

**Answer:** C

**Explanation:**

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

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

**NEW QUESTION 125**

- (Exam Topic 3)

You are producing a multiple linear regression model in Azure Machine Learning Studio. Several independent variables are highly correlated.

You need to select appropriate methods for conducting elective feature engineering on all the data.

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

- A. Mastered
- B. Not Mastered

**Answer:** A

**Explanation:**

**NEW QUESTION 130**

- (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 133

- (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.

- 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 135

- (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.

- 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 140**

- (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 143**

.....

## **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](#)**