- **Vendor: Microsoft**
- **Exam Code: DP-100**
- **Exam Name: Designing and Implementing a Data Science Solution on Azure**
- **New Updated Questions from Braindump2go (Updated in May/2020)**



**QUESTION 94**
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:** ABE
**Explanation:**
AE: The algorithm terminates when the centroids stabilize or when a specified number of iterations are completed.
B: A measure of how well the centroids represent the members of their clusters is the residual sum of squares or RSS, the squared distance of each vector from its centroid summed over all vectors. RSS is the objective function and our goal is to minimize it.
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

**QUESTION 95**
You design data engineering solutions for a company.
A project requires analytics and visualization of large set of data. The project has the following requirements:
- Notebook scheduling
- Cluster automation
- Power BI Visualization

You need to recommend the appropriate Azure service. Which Azure service should you recommend?

A. Azure Batch
B. Azure Stream Analytics
C. Azure ML Studio
D. Azure Databricks
E. Azure HDInsight

**Answer:** D

**QUESTION 96**
You are developing a solution that performs real-time analysis of IoT data in the cloud. The solution must remain available during Azure service updates.

You need to recommend a solution.
Which two actions should you recommend? Each correct answer presents part of the solution. NOTE: Each correct selection is worth one point.

A. Deploy an Azure Stream Analytics job to two separate regions that are not in a pair.
B. Deploy an Azure Stream Analytics job to each region in a paired region.
C. Monitor jobs in both regions for failure.
D. Monitor jobs in the primary region for failure.
E. Deploy an Azure Stream Analytics job to one region in a paired region.

**Answer:** BC

**QUESTION 97**
A company is designing a solution that uses Azure Databricks.
The solution must be resilient to regional Azure datacenter outages.
You need to recommend the redundancy type for the solution.
What should you recommend?

A. Read-access geo-redundant storage
B. Locally-redundant storage
C. Geo-redundant storage
D. Zone-redundant storage

**Answer:** C

**QUESTION 98**
A company has a Microsoft Azure HDInsight solution that uses different cluster types to process and analyze data.
Operations are continuous. Reports indicate slowdowns during a specific time window.
You need to determine a monitoring solution to track down the issue in the least amount of time.
What should you use?

A. Azure Log Analytics log search query
B. Ambari REST API
C. Azure Monitor Metrics
D. HDInsight .NET SDK
E. Azure Log Analytics alert rule query

**Answer:** B

**QUESTION 99**
You are designing a solution for a company. The solution will use model training for objective classification.
You need to design the solution.
What should you recommend?

A. an Azure Cognitive Services application
B. a Spark Streaming job
C. interactive Spark queries
D. Power BI models
E. a Spark application that uses Spark MLib.

**Answer:** E

**QUESTION 100**
A company manufactures automobile parts. The company installs IoT sensors on manufacturing machinery.
You must design a solution that analyzes data from the sensors.
You need to recommend a solution that meets the following requirements:

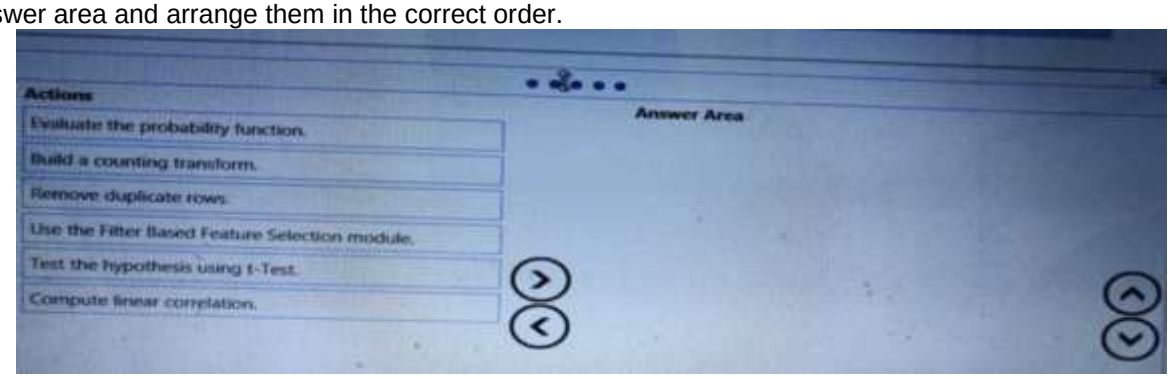
- Data must be analyzed in real-time.
- Data queries must be deployed using continuous integration.
- Data must be visualized by using charts and graphs.
- Data must be available for ETL operations in the future.
- The solution must support high-volume data ingestion.

Which three actions should you recommend? Each correct answer presents part of the solution. NOTE: Each correct selection is worth one point.

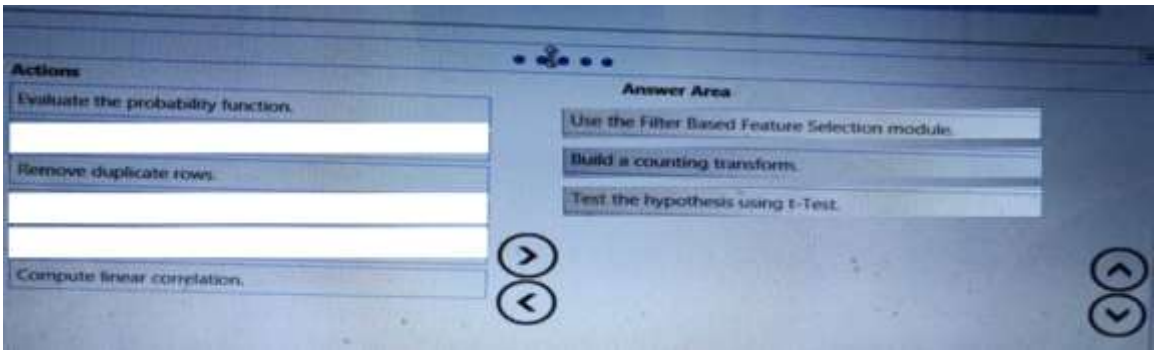
A. Use Azure Analysis Services to query the data. Output query results to Power BI.
B. Configure an Azure Event Hub to capture data to Azure Data Lake Storage.
C. Develop an Azure Stream Analytics application that queries the data and outputs to Power BI. Use Azure Data Factory to deploy the Azure Stream Analytics application.
D. Develop an application that sends the IoT data to an Azure Event Hub.
E. Develop an Azure Stream Analytics application that queries the data and outputs to Power BI. Use Azure Pipelines to deploy the Azure Stream Analytics application.
F. Develop an application that sends the IoT data to an Azure Data Lake Storage container.

**Answer:** BCD

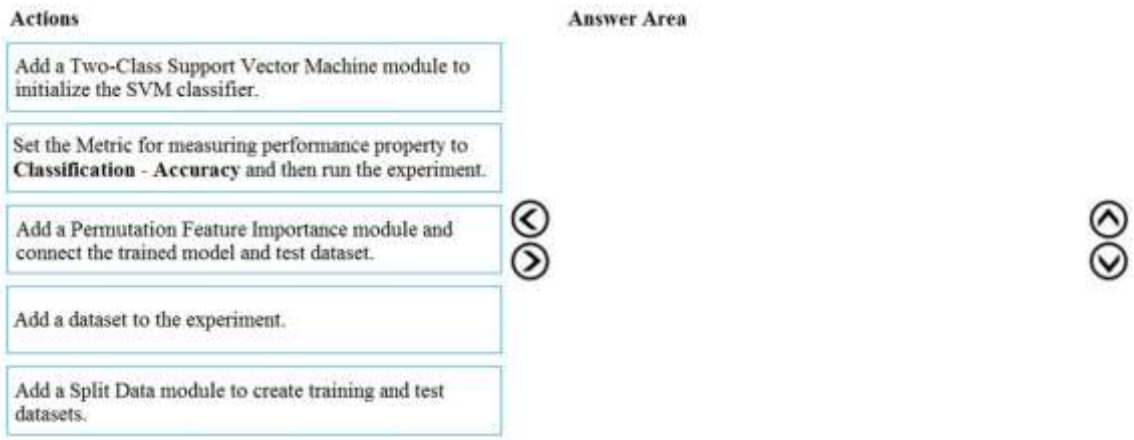
**QUESTION 101**
Drag and Drop Question
You are producing a multiple linear regression model in Azure Machine learning Studio.
Several independent variables are highly correlated.
You need to select appropriate methods for conducting elective feature engineering on all the data.
Which three actions should 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.



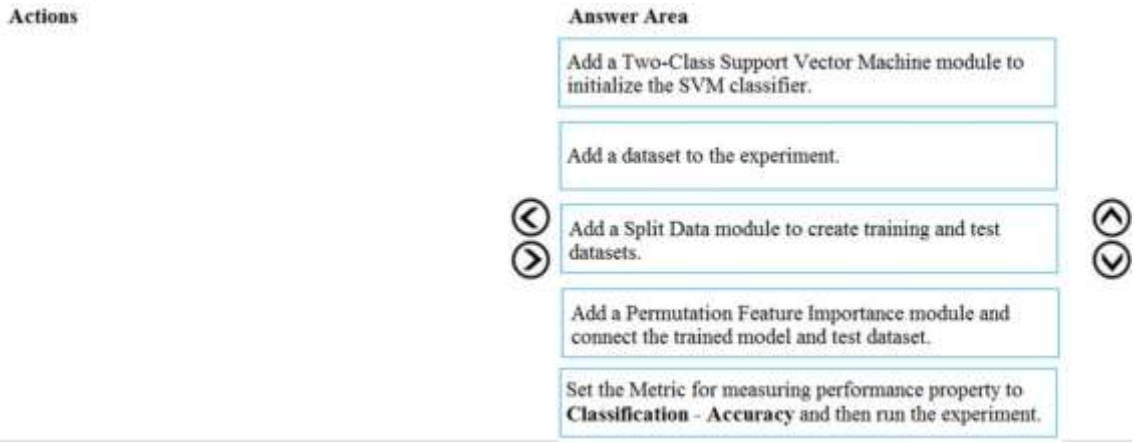

**Answer:**




**QUESTION 102**
Drag and Drop Question
YOU have a data-set that contains over 150 features. You use the dataset to train a Support Vector Machine (SVM) binary classifirer.
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 al actions from from the list of Actions to the answer area and arrange them in the correct order.




**Answer:**




**QUESTION 103**
Hotspot Question
You are creating a machine learning model in Python. The provided dataset contains several numerical columns and one text column.

- `Biker`
- `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.

**Answer Area**

```
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)
```

**Answer:**

**Answer Area**

```
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)
```

**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/

**QUESTION 104**
Hotspot Question
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.

**Answer Area**

| Description | Error type |
|---|---|
| A person has a disease. The model classifies the case as having a disease. | True Positives / True Negatives / False Positives / False Negatives |
| A person does not have a disease. The model classifies the case as having no disease. | True Positives / True Negatives / False Positives / False Negatives |
| A person does not have a disease. The model classifies the case as having a disease. | True Positives / True Negatives / False Positives / False Negatives |
| A person has a disease. The model classifies the case as having no disease. | True Positives / True Negatives / False Positives / False Negatives |

**Answer:**

| Description | Error type |
| --- | --- |
| A person has a disease. The model classifies the case as having a disease. | True Positives |
| A person does not have a disease. The model classifies the case as having no disease. | True Negatives |
| A person does not have a disease. The model classifies the case as having a disease. | False Positives |
| A person has a disease. The model classifies the case as having no disease. | False Negatives |

**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