

PRACTICE QUESTIONS

Databricks Certified Machine Learning Associate

SHARED BY [u/Witty-Pebble-6781](#) on aiexamcerts

EDITION Free Set 2

QUESTIONS 10

SHARED October 15, 2026

EXAM Databricks Certified Machine Learning Associate

Shared by [u/Witty-Pebble-6781](#) on aiexamcerts

10 practice questions for the Databricks Certified Machine Learning Associate exam, each followed by its answer and a short explanation. More free sets and the full practice exam:

www.aiexamcerts.com/databricks/databricks-certified-machine-learning-associate. Spotted a mistake? Email support@aiexamcerts.com with the exam name and the question number.

1. Where are registered models governed and shared across workspaces in Databricks?

- A The DBFS root of a single workspace.
- B Unity Catalog under a catalog.schema.model name.
- C The legacy workspace model registry in each workspace.
- D A Git folder in one workspace.

Answer: B

Unity Catalog stores registered models under the catalog and schema hierarchy, so access control and sharing apply across workspaces. The legacy workspace registry, the DBFS root and a Git folder are each tied to a single workspace, and the legacy registry is the tempting choice.

2. A feature table's primary key must be unique for each row, so that a lookup returns exactly one set of feature values for an entity.

- A True
- B False

Answer: A (True)

The primary key identifies one row per entity, which makes point lookups deterministic when training sets or online features are assembled. Duplicate keys would make the join ambiguous. The statement is true.

3. A team wants colleagues to re-run a training job with the same code, environment and entry point. Which MLflow component packages the project for this purpose?

- A The MLflow Tracking UI
- B MLflow Model Serving
- C Autoscaling cluster policies
- D MLflow Projects

Answer: D

MLflow Projects defines the code, dependencies and entry points so a run can be reproduced elsewhere. The Tracking UI records and displays runs but does not package the code that produced them, which makes it the tempting choice. Model Serving and cluster policies address deployment and compute rather than project packaging.

4. A Spark function applied to batches of rows as pandas data for vectorized processing is called a pandas _____.

Answer: UDF (also accepted: user-defined function)

A pandas UDF processes columns in batches using pandas and Arrow, which is faster than a row-at-a-time Python UDF. Built-in Spark SQL functions are the first choice when they exist. The blank refers to the vectorized user-defined function type.

5. Which two statements about comparing MLflow runs are correct? (Choose two.)

- A Runs can be sorted and filtered by metric values in the experiment view.
- B Two runs with the same parameters always produce identical metric values on GPUs.
- C Logged parameters can be edited after the run finishes.
- D Metrics logged with step values can be plotted over the training iterations.
- E A registered model version automatically compares itself with every run.

Answer: A, D

The experiment view supports sorting and filtering by metrics, and metrics logged with steps appear as curves over training. Same parameters do not guarantee identical GPU results, and logged parameters are fixed once written. Registration does not run a comparison against every run.

6. A dataset fits comfortably in memory on a single-node cluster, and the team prefers scikit-learn with pandas. Which approach is most appropriate?

- A Use SparkTrials for every classical model.
- B Use scikit-learn directly on the pandas DataFrame.
- C Rewrite the model as a Spark ML pipeline before training.
- D Train the model on a SQL warehouse.

Answer: B

When the data fits in memory, scikit-learn on pandas is simpler and well supported, and Spark distribution adds overhead without benefit. SparkTrials distributes hyperparameter trials across a cluster, which is the tempting but unnecessary choice here. Rewriting the model as a Spark ML pipeline adds overhead for no benefit, and a SQL warehouse does not train models.

7. What does the Spark ML BinaryClassificationEvaluator compute by default?

- A The mean absolute error.
- B The silhouette score.
- C The within-cluster sum of squares.
- D The area under the ROC curve.

Answer: D

The default metric of BinaryClassificationEvaluator is the area under the ROC curve, and the precision-recall variant is available through the metric name option. Mean absolute error belongs to regression evaluation, which is the tempting mismatch. Silhouette and within-cluster sum of squares describe clustering.

8. The test set should be used repeatedly during hyperparameter tuning so that the final estimate reflects the best configuration.

- A True
- B False

Answer: B (False)

Repeated use of the test set for selection turns it into a second validation set, so its score becomes optimistic. Tuning should use a validation set or cross-validation, and the test set should be reserved for one final estimate. The statement is false.

9. Which tuning method samples settings at random from a search space within a fixed budget of trials?

- A Forward feature selection
- B Principal component analysis
- C Target encoding
- D Random search

Answer: D

Random search draws configurations from the space, which often covers important hyperparameters better than a grid of the same size. Forward feature selection chooses inputs rather than settings, which is the tempting confusion. PCA and target encoding are preprocessing methods.

10. Which setting makes a random hyperparameter search reproducible across runs?

- A A fixed random seed for the sampler.
- B The maximum possible Hyperopt parallelism.
- C A new experiment name for each run.
- D A larger evaluation budget.

Answer: A

Fixing the seed makes the sampler draw the same configurations on each run, so results can be repeated and compared. Maximum parallelism changes the order in which results arrive, which is the tempting but insufficient fix. Experiment names and a larger evaluation budget do not control the sampling.