

PRACTICE QUESTIONS

DASCA

Principal Data Scientist

SHARED BY [u/curious_seal_27](#) on aiexamcerts

EDITION Free Set 1

QUESTIONS 10

SHARED October 15, 2026

EXAM DASCA Principal Data Scientist

Shared by u/curious_seal_27 on aiexamcerts

10 practice questions for the DASCA Principal Data Scientist exam, each followed by its answer and a short explanation. More free sets and the full practice exam: www.aiexamcerts.com/dasca/principal-data-scientist. Spotted a mistake? Email support@aiexamcerts.com with the exam name and the question number.

1. Which design best estimates the causal effect of a new recommendation feature on purchases?
 - A Comparing purchases before and after launch for all users with no control group at all
 - B Randomly assigning users to see the feature or not, then comparing their outcomes
 - C Asking the most active users in a survey whether they like the feature or not
 - D Choosing users who clicked the feature and comparing them with those who did not click

Answer: B

Random assignment makes the two groups comparable on average, so a difference in purchases can be attributed to the feature. Before-and-after comparisons absorb seasonality, and comparing clickers with non-clickers or surveying active users introduces selection bias.

2. A version-controlled registry that tracks trained model artifacts, their metadata, and their lifecycle stage, such as staging or production, is called a model _____.

Answer: registry (also accepted: model registry)

A model registry records each version with its metadata and stage, so teams know which artifact serves production and can roll back to a previous one. It is distinct from a feature store, which manages the inputs to models.

3. When is a bootstrap confidence interval especially useful?
 - A When the data are guaranteed to be perfectly normal in every dimension measured
 - B When the dataset contains exactly one observation in each of its categories and groups
 - C When the analyst already knows the exact population parameter in advance of the study
 - D When a metric lacks a simple standard error or its assumptions are hard to check

Answer: D

Bootstrapping resamples the observed data to approximate the sampling distribution of almost any statistic, which is valuable when no closed-form standard error exists. It is not needed when the parameter is already known, and it does not depend on data being perfectly normal.

4. A feature store helps teams reuse consistent feature definitions between model training and serving.

- A True
- B False

Answer: A (True)

A feature store centralizes feature definitions and computation so that training and online serving use the same logic, which reduces training-serving skew. The statement is true.

5. Which two factors most strongly determine whether an enterprise can scale machine learning beyond pilots? (Choose two.)

- A Reliable data pipelines and clear ownership of production data
- B The number of business units that have attended an introductory analytics workshop
- C Repeatable deployment and monitoring processes for models
- D The programming language that most data scientists in the company prefer to use
- E The number of proof-of-concept demos presented during the pilot phase

Answer: A, C

Production models depend on dependable data pipelines with clear owners, and on repeatable deployment and monitoring. Workshop attendance, language preference, and demo counts do not determine whether a model can operate reliably at scale.

6. Which approach addresses class imbalance without changing the evaluation data?

- A Deleting all minority examples so that the classes match in size
- B Oversampling the minority class in the test set before scoring the model
- C Oversampling the minority class in training folds only
- D Reporting only accuracy because imbalance no longer matters for scoring

Answer: C

Resampling only the training folds rebalances what the model learns from while the validation and test data keep their real distribution. Oversampling the test set distorts the evaluation, deleting minority records discards information, and reporting accuracy alone hides the imbalance.

7. Which architecture was most commonly used for sequence modeling tasks such as translation before transformers dominated the field?
- A Principal component analysis applied to the input vocabulary of the corpus
 - B Convolutional networks used only for static image classification problems
 - C Recurrent neural networks, including LSTM variants
 - D Hierarchical clustering with a fixed number of levels in the dendrogram

Answer: C

Recurrent networks, including LSTM variants, processed sequences step by step and were the standard for translation and language modeling before transformers. Convolutional networks for static images, PCA, and hierarchical clustering are not sequence models.

8. The property that lets a pipeline step be rerun safely, producing the same result without duplicating or corrupting data, is called _____.

Answer: idempotence (also accepted: idempotent, idempotency)

Idempotent steps can be retried after a failure without side effects, which is essential for scheduled data pipelines that will sometimes rerun. Without it, a retry can double-count records or overwrite good results with partial ones.

9. A stakeholder rejects a model because it is a black box. Which response best addresses the concern?
- A Replace the model with a rule list chosen without any analysis of the data
 - B Explain that all state-of-the-art models are black boxes and ask them to trust the vendor
 - C Show global and local explanations plus validation evidence for the decision
 - D Shorten the presentation to skip the technical details entirely for the group

Answer: C

Global explanations show what drives the model overall, local explanations show why an individual decision was made, and validation evidence shows it works. Asking for trust, replacing the model without analysis, or skipping details does not answer the concern.

10. Under the GDPR, which principle requires that personal data be kept only as long as necessary for the stated purpose?
- A Right of access
 - B Data portability
 - C Storage limitation
 - D Consent withdrawal

Answer: C

Storage limitation is a core GDPR principle requiring that personal data not be kept longer than needed for its purpose. Portability, access, and consent withdrawal are data subject rights rather than the retention principle.