

PRACTICE QUESTIONS

Google Cloud

Professional Machine Learning Engineer

SHARED BY [u/smoothllama1993](#) on aiexamcerts

EDITION Free Set 2

QUESTIONS 10

SHARED October 15, 2026

EXAM Google Cloud Professional Machine Learning Engineer

Shared by [u/smoothllama1993](#) on aiexamcerts

10 practice questions for the Google Cloud Professional Machine Learning Engineer exam, each followed by its answer and a short explanation. More free sets and the full practice exam:

www.aiexamcerts.com/google-cloud/professional-machine-learning-engineer. Spotted a mistake? Email support@aiexamcerts.com with the exam name and the question number.

1. A data science team needs a Google-managed notebook environment integrated with Vertex AI for exploratory work. Which service fits?

- A Cloud Run
- B Cloud Composer
- C A self-managed Compute Engine VM with Jupyter installed by hand
- D Vertex AI Workbench

Answer: D

Vertex AI Workbench provides managed notebooks that connect to Vertex AI data and training services. A hand-installed VM is the tempting option but leaves the team to manage it. Cloud Run serves containers and Cloud Composer orchestrates workflows, so neither is a notebook environment.

2. A Vertex AI pipeline can be scheduled to run recurrently, such as every week, without someone starting each run by hand.

- A True
- B False

Answer: A (True)

The statement is true. Pipeline runs can be scheduled, so retraining or scoring happens automatically on a cadence. Manual starts are only needed for ad hoc runs.

3. A loan model approves applicants from one demographic group at a much lower rate than another. Which action is the most appropriate first step?

- A Ignoring the gap because overall accuracy is high
- B Evaluating fairness metrics by group, investigating feature and data causes, and involving the governance owner before release
- C Removing the group column from the report only
- D Lowering the decision threshold for everyone to hide the gap

Answer: B

Measuring outcomes by group, finding the causes and bringing in the accountable owner is the responsible sequence before any release decision. High overall accuracy can coexist with unfair group outcomes. Hiding the column or shifting thresholds conceals the problem without fixing it.

4. To absorb variable request volume, an endpoint's autoscaling range is set with minimum and maximum _____ counts.

Answer: replica (also accepted: replica count)

Replicas are the serving instances that the endpoint scales between its minimum and maximum. The minimum keeps capacity ready, and the maximum caps cost. The blank is replica.

5. Which two practices help teams collaborate on ML models safely? (Choose two.)

- A Versioning training code and model artifacts, with metadata linking them
- B Editing the production model directly from personal notebooks
- C Granting least-privilege IAM roles on data and model resources
- D Sharing service account keys in a team chat channel
- E Keeping experiment results only in personal files

Answer: A, C

Versioned code and artifacts with linked metadata make results reproducible, and least-privilege roles limit who can change data or models. Editing production from personal notebooks and sharing keys in chat are both unsafe. Personal-only results cannot be reviewed by the team.

6. Which storage format is commonly recommended for large training datasets read by TensorFlow input pipelines on Google Cloud?

- A Plain text lines with no schema or typing
- B Excel workbooks
- C TFRecord files read with tf.data
- D Screenshots of tables

Answer: C

TFRecord stores serialized examples in a compact, streamable format that tf.data reads efficiently in parallel. Plain text lacks a schema and is slower to parse. Spreadsheets and screenshots are not designed for training input at scale.

7. Which property matters most when an online model serves a mobile app that sees sudden traffic bursts?

- A How long the training job took
- B The number of notebooks in the project workspace
- C The name of the model
- D Autoscaling and latency under peak load

Answer: D

Users experience bursts through latency and availability, so the endpoint must scale and hold its response time at peak. Training duration, notebook counts and model names do not affect the serving experience during a burst.

8. It is acceptable to split a time-series dataset randomly into training and test sets when the goal is to forecast future values.
- A True
 - B False

Answer: B (False)

The statement is false. A random split lets the model train on later observations than some test points, which leaks future information. A time-ordered split matches how the model will be used.

9. Which Vertex AI capability records pipeline runs and the artifacts each step produced, so lineage can be traced?
- A Cloud Logging sinks only
 - B Vertex ML Metadata
 - C Cloud DNS records
 - D BigQuery reservations

Answer: B

Vertex ML Metadata stores the lineage of executions, artifacts and parameters across pipeline runs. Logging sinks export log lines, but they do not model artifact lineage. DNS records and BigQuery reservations are unrelated to ML lineage.

10. A model's accuracy dropped after a market change, although the input schema has not changed. What is the most likely cause?
- A A hardware failure in the training cluster
 - B A missing TensorBoard setting
 - C Concept drift, where the relationship between inputs and the target has changed
 - D A larger training batch size configured in the next scheduled retraining job on the cluster

Answer: C

When the inputs look the same but the outcomes behave differently, the underlying relationship has shifted, which is concept drift. A schema change would show up as a validation error rather than a quiet accuracy drop. Hardware, visualization settings and batch size do not change what the model has learned.