

PRACTICE QUESTIONS

CompTIA DataAI

SHARED BY [u/Curious-Hawk-9121](#) on aiexamcerts

EDITION	Free Set 2
QUESTIONS	10
SHARED	October 15, 2026
EXAM	CompTIA DataAI

Shared by u/Curious-Hawk-9121 on aiexamcerts

10 practice questions for the CompTIA DataAI exam, each followed by its answer and a short explanation. More free sets and the full practice exam: www.aiexamcerts.com/comptia/comptia-dataai. Spotted a mistake? Email support@aiexamcerts.com with the exam name and the question number.

1. A two-sided hypothesis test returns a p-value of 0.03, and the analyst set the significance level at 0.05 before running the test. What is the correct decision?
 - A Fail to reject the null hypothesis, because 0.03 is above 0.05.
 - B Reject the null hypothesis, because the p-value is below the 0.05 significance level.
 - C Prove the alternative hypothesis is true, because the chance that the null is wrong is 97 percent.
 - D Conclude that the sample was too small for the result to matter.

Answer: B

When the p-value is less than the chosen significance level, the null hypothesis is rejected. The p-value is not the probability that the null hypothesis is true, and it does not prove the alternative, so reading it as 97 percent certainty is the most tempting misreading. Reversing the comparison and ignoring the test result entirely are the other wrong readings.

2. Bayes' theorem states that $P(A \text{ given } B)$ always equals $P(B \text{ given } A)$.
 - A True
 - B False

Answer: B (False)

Bayes' theorem relates the two conditional probabilities through the prior probabilities of A and B, so they are generally different. Treating them as equal is the classic prosecutor's fallacy and produces badly wrong posterior estimates. The statement is false.

3. A fraud model reports 99 percent accuracy on a dataset where only 1 percent of transactions are fraudulent. What is the most likely problem with trusting that figure?
 - A The model is underfitting the training data, so it needs a simpler architecture.
 - B The learning rate was set too high.
 - C The model needs additional hidden layers.
 - D Accuracy is misleading under heavy class imbalance, so precision and recall should be examined.

Answer: D

A model that labels every transaction as legitimate would score 99 percent here while catching no fraud, so accuracy hides the failure. Precision and recall for the fraud class expose it. Underfitting, learning rate and network depth are plausible explanations for other problems, but they do not explain why this accuracy number is uninformative.

4. The count of actual positive cases that the model correctly labels as positive is called the number of true _____.

Answer: positives

True positives are one cell of the confusion matrix, alongside true negatives, false positives and false negatives. Precision and recall are both computed from this count. The term refers to correctly predicted positive cases, not to the total number of positives in the data.

5. Which two statements about the k-nearest neighbors algorithm are correct? (Choose two.)

- A It builds an explicit parametric model with a fixed set of learned weights.
- B Its prediction time is constant regardless of the size of the training set.
- C It needs no feature scaling because distances are never computed.
- D Its accuracy can degrade with many irrelevant features because distances become less informative.
- E A very small value of k can make predictions sensitive to noise.

Answer: D, E

Distances in high dimensions are dominated by irrelevant features, which weakens the neighbor relationship, and a very small k follows individual noisy points. KNN is non-parametric and stores the training data, so it has no fixed weights and its prediction cost grows with the training set. It also depends on distances, so feature scaling matters.

6. In a convolutional neural network, what is the primary role of a pooling layer?

- A To learn new color channels from the input image.
- B To downsample feature maps, reducing spatial size and computation.
- C To normalize the text tokens before embedding.
- D To output the final class probabilities.

Answer: B

Pooling, such as max pooling, summarizes small neighborhoods of a feature map, which shrinks the spatial dimensions and reduces computation while keeping the strongest responses. Learning features is the job of the convolution filters, which makes learning new color channels a tempting confusion. Output probabilities usually come from a softmax layer, not from pooling.

7. What is a data warehouse optimized for?

- A High-volume individual transactional writes, with many small inserts and updates per second.
- B Storing unstructured images only.
- C Analytical queries and reporting over integrated historical data.
- D Real-time caching of web sessions.

Answer: C

A warehouse integrates data from several sources into subject-oriented, historical tables that are optimized for aggregation and reporting. Transactional systems are the tempting confusion because they hold similar business data, but they are optimized for many small writes. Image storage and session caching are different workloads.

8. A canary deployment sends all production traffic to the new model version at once, so it does not need monitoring.

- A True
- B False

Answer: B (False)

A canary deployment routes a small share of traffic to the new version first and compares its behavior with the current version before expanding. Sending everything at once is a full cutover, which removes the safety net that the canary provides. The statement is false.

9. Which statement about anomaly detection is most accurate?

- A It always requires a large labeled set of anomalies before any model can be trained.
- B It cannot be applied to numeric sensor data.
- C It can use methods such as isolation forests or z-score thresholds when anomalies are rare or unlabeled.
- D It is the same task as multiclass classification.

Answer: C

Unsupervised and semi-supervised methods, such as isolation forests or simple statistical thresholds, flag unusual observations without many labeled anomalies. Requiring a large labeled anomaly set is the most tempting misconception, because supervised classification does need one. Numeric sensor readings are a common anomaly detection input.

10. What does a fairness metric such as demographic parity check?
- A Whether training finished within the time budget.
 - B Whether positive prediction rates are similar across protected groups.
 - C Whether the model's overall accuracy is better than a coin toss on the holdout set.
 - D Whether the dataset has no missing values.

Answer: B

Demographic parity compares the rate of positive predictions between groups, so a large gap signals a disparity that may need investigation. Overall accuracy can look good while one group is treated worse, which makes accuracy a tempting but insufficient check. Training time and missing values are unrelated to this metric.