

PRACTICE QUESTIONS

CertNexus

DSP-210

Certified Data Science Practitioner

SHARED BY [u/mellowotter1999](#) on aiexamcerts

EDITION Free Set 2

QUESTIONS 10

SHARED October 15, 2026

EXAM CertNexus DSP-210, Certified Data Science Practitioner

Shared by [u/mellowotter1999](#) on aiexamcerts

10 practice questions for the CertNexus DSP-210 exam, Certified Data Science Practitioner, each followed by its answer and a short explanation. More free sets and the full practice exam: www.aiexamcerts.com/certnexus/dsp-210. Spotted a mistake? Email support@aiexamcerts.com with the exam code DSP-210 and the question number.

1. Which algorithm family is most naturally used to group unlabeled customers into segments?

- A Linear regression with a continuous target variable and many predictors
- B Support vector classification trained on labeled churn outcomes for customers
- C Time series forecasting of monthly revenue totals for each product line
- D Clustering such as k-means on unlabeled customer features

Answer: D

Clustering groups records by similarity without needing labels, which is exactly what unlabeled customer segmentation requires. The other families need a target variable or a time-ordered series, so they do not fit an unlabeled segmentation task.

2. Standardization rescales each numeric feature to a mean of zero and a standard deviation of one.

- A True
- B False

Answer: A (True)

Standardization subtracts the feature mean and divides by its standard deviation, producing zero mean and unit variance. Scikit-learn implements it with StandardScaler. The statement is true.

3. A data scientist needs to attach customer region data to an orders table using a shared customer_id column. Which pandas call fits this need?

- A DataFrame.drop_duplicates on the order_id column
- B DataFrame.describe for summary statistics
- C DataFrame.to_parquet for file export
- D DataFrame.merge on customer_id

Answer: D

DataFrame.merge joins two tables on a shared key, which is what attaching region data to orders requires. Describe summarizes values, to_parquet writes files, and drop_duplicates removes rows, so none of them combines two tables.

4. Performing k-fold cross-validation with k equal to 5 means the data is split into _____ folds, and each fold is used once as the validation set.

Answer: five (also accepted: 5)

With five folds, each part serves once as the validation set while the remaining four train the model, and the scores are averaged. This gives a more stable estimate than one split, which depends on a single partition.

5. Which two steps help prevent a machine learning pipeline from leaking information from the test set? (Choose two.)
- A Fitting imputers and scalers on the full dataset before splitting it
 - B Tuning hyperparameters repeatedly against the final test score
 - C Putting preprocessing inside a Pipeline that is fit only on training folds
 - D Holding the test split out until final evaluation, after all choices are fixed
 - E Reusing the test split as the validation set during model selection

Answer: C, D

A pipeline fit only on training folds keeps preprocessing statistics away from held-out data, and holding out the test split until the end keeps it an honest measure. Fitting on the full dataset, tuning against the test score, and reusing the test split for selection all leak information.

6. Which metric summarizes a binary classifier's ability to rank positives above negatives across all thresholds?
- A Silhouette score
 - B Mean squared error
 - C ROC AUC
 - D Inertia of the cluster centers

Answer: C

ROC AUC is the area under the receiver operating characteristic curve and measures ranking quality across every threshold. Mean squared error is a regression measure, while silhouette score and inertia evaluate clustering rather than a classifier.

7. A grouped comparison shows one trend, but the combined data shows the opposite trend. What is this called?
- A Multicollinearity
 - B Regression to the mean
 - C Survivorship bias
 - D Simpson's paradox

Answer: D

Simpson's paradox occurs when a lurking grouping variable reverses a trend that appears within each group, so the aggregate and subgroup views disagree. Regression to the mean, survivorship bias, and multicollinearity describe different statistical effects.

8. A model card or datasheet can help document the limitations and intended uses of a model.
- A True
 - B False

Answer: A (True)

Model cards and datasheets record intended uses, evaluation context, data sources, and known limitations, so downstream users can judge fit. The statement is true.

9. Before deploying a churn model, what should the team document for reproducibility?
- A Only the final accuracy number from the single best experiment run
 - B The data version, code version, random seeds, and package versions used
 - C The name of the person who wrote the first cell of the initial notebook draft
 - D The slide template used for the final stakeholder review

Answer: B

Reproducing a model requires the exact data snapshot, code, seeds, and library versions, so anyone can rebuild the same result. A single accuracy number or the author's name does not allow the work to be rerun.

10. Which visualization best compares the revenue of five product lines at one point in time?
- A A heat map of GPS coordinates for every delivery
 - B A line chart spanning twenty years of daily timestamps and their values
 - C A network graph of supplier relationships across the entire supply chain
 - D A bar chart sorted by revenue

Answer: D

A sorted bar chart makes the ranking and size of each product line easy to compare at a single point in time. A long line chart suits trends over time, while network graphs and GPS heat maps answer other questions.