
DUMPSGATE
98.99%
PASS RATE
SUCCESS
100%
GUARANTEED
UNLIMITED
FREE UPDATES
REAL EXAM
CERTIFICATION DUMPS
PDF
INSTANT PDF DOWNLOAD
REAL EXAM QUESTIONS
AUTHENTIC ANSWERS

A CROSS JOIN produces the Cartesian product of the two tables (every row from the first paired with every row from the second), yielding far more rows than any of the other join types.

Question: 30

A data scientist built several models that perform about the same but vary in the number of features. Which of the following models should the data scientist recommend for production according to Occam's razor?

A. The model with the fewest features and highest performance

B. The model with the fewest features and the lowest performance

C. The model with the most features and the lowest performance

D. The model with the most features and the highest performance

Answer: A

Explanation:

According to Occam’s razor, when models perform equivalently, you choose the simplest one - in this case, the model that achieves the needed performance with the fewest features.

Question: 31

A data analyst wants to use compression on an analyzed data set and send it to a new destination for further processing. Which of the following issues will most likely occur?

A. Library dependency will be missing.

B. Server CPU usage will be too high.

C. Operating system support will be missing.

D. Server memory usage will be too high.

Answer: B

Explanation:

Compression and decompression are CPU-intensive operations; on large data sets, the extra processing load can significantly spike CPU utilization. Memory, OS support, or library dependencies are far less likely to be the primary bottleneck in a standard compression workflow.

Question: 32

The most likely concern with a one-feature, machine-learning model is high error due to:

A. bias

B. dimensionality.

C. variance.

D. probability.

Answer: A

Explanation:

A model with only one feature is unlikely to capture the true complexity of the data's underlying relationships, leading to systematic underfitting - i.e., high bias.

Question: 33

Given matrix

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \\ 3 & 2 & 1 \end{bmatrix}$$

Which of the following is AT?

A)

$$\begin{bmatrix} 3 & 2 & 1 \\ 2 & 1 & 3 \\ 1 & 2 & 3 \end{bmatrix}$$

B)

$$\begin{bmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \\ 3 & 2 & 1 \end{bmatrix}$$

C)

$$\begin{bmatrix} 1 & 2 & 3 \\ 2 & 1 & 2 \\ 3 & 3 & 1 \end{bmatrix}$$

D)

$$\begin{bmatrix} 3 & 3 & 1 \\ 2 & 1 & 2 \\ 1 & 2 & 3 \end{bmatrix}$$

A. Option A

B. Option B

C. Option C

D. Option D

Answer: C

Explanation:

Transposing swaps rows and columns, so the (i, j) entry becomes the (j, i) entry.

Question: 34

A data scientist is clustering a data set but does not want to specify the number of clusters present. Which of the following algorithms should the data scientist use?

A. DBSCAN

B. k-nearest neighbors

C. k-means

D. Logistic regression

Answer: A

Explanation:

DBSCAN discovers clusters based on density without requiring you to predefine the number of clusters, automatically finding arbitrarily shaped groups and identifying noise points.

Question: 35

A data analyst wants to find the latitude and longitude of a mailing address. Which of the following is the best method to use?

A. One-hot encoding

B. Binning

C. Geocoding

D. Imputing

Answer: C

Explanation:

Geocoding is the process of converting a postal address into geographic coordinates (latitude and longitude), making it the appropriate method.

Question: 36

Which of the following describes the appropriate use case for PCA?

A. Dimensionality reduction

B. Classification

C. Regression

D. Recommendation

Answer: A

Explanation:

Principal Component Analysis transforms correlated features into a smaller set of uncorrelated components that capture most of the variance, making it ideal for reducing dimensionality before modeling or visualization.

Question: 37

A data scientist observes findings that indicate that as electrical grids in a country become more and more connected over time, the frequency of brownouts and blackouts in total decrease, and the frequency of major brownouts and blackouts increase. Which of the following distribution metrics could best be identified?

A. Scale axis magnitudes

B. Kurtosis

C. Skewness

D. Normality

Answer: B

Explanation:

Kurtosis quantifies how heavy or light the tails of a distribution are. In this case, fewer overall events but more extreme (major) brownouts/blackouts indicates heavier tails over time. This is exactly what an increasing kurtosis would reveal.

Question: 38

A data scientist is merging two tables. Table 1 contains employee IDs and roles. Table 2 contains employee IDs and team assignments. Which of the following is the best technique to combine these data sets?

A. inner join between Table 1 and Table 2

B. left join on Table 1 with Table 2

C. right join on Table 1 with Table 2

D. outer join between Table 1 and Table 2

Answer: A

Explanation:

An INNER JOIN merges records only where the employee ID exists in both tables, yielding a single combined table of each employee's role paired with their team assignment.

Question: 39

Which of the following is a classic example of a constrained optimization problem?

A. The cold start problem

B. The traveling salesman

C. Calculating local maximum

D. Calculating gradient descent

Answer: B

Explanation:

The traveling-salesman problem seeks the shortest possible route that visits each city exactly once and returns to the start, making it a textbook example of optimization under explicit constraints.

Question: 40

A data scientist wants to digitize historical hard copies of documents. Which of the following is the best method for this task?

A. Word2vec

B. Optical character recognition

C. Latent semantic analysis

D. Semantic segmentation

Answer: B

Explanation:

OCR converts scanned images of text into machine-readable characters, making it the appropriate tool for digitizing printed or handwritten historical documents.

Question: 41

A data scientist trained a model for departments to share. The departments must access the model using HTTP requests. Which of the following approaches is appropriate?

A. Utilize distributed computing.

B. Deploy containers.

C. Create an endpoint.

D. Use the File Transfer Protocol.

Answer: C

Explanation:

Exposing the model behind an HTTP endpoint (for example, a REST API) allows other departments to send requests and receive predictions directly over HTTP. The other options don’t inherently provide a request–response interface for sharing a model.

Question: 42

Given the following:

$$X_t = \delta + \phi_1 X_{t-1} + \omega_t \text{ where } \omega_t \sim N(0, \sigma_\omega^2)$$

Which of the following time series models best represents this process?

A. ARIMA(1,1,1)

B. ARMA(1,1)

C. SARIMA(1, 1, 1) x (1, 1, 1)1

D. AR(1)

Answer: D

Explanation:

The model has a single autoregressive term and only white-noise errors, matching the definition of an AR(1) process.

Question: 43

Which of the following methods should a data scientist use just before switching to a potential replacement model?

A. A/B testing

B. Performance monitoring

C. CI/CD

D. Containerization

Answer: A

Explanation:

A/B testing lets you compare the current model against the candidate in parallel, measuring performance on live data, before fully switching to the new model.

Question: 44

A data scientist is presenting the recommendations from a monthslong modeling and experiment process to the company's Chief Executive Officer. Which of the following is the best set of artifacts to include in the presentation?

A. Methods, data overview, results, recommendations, and charts

B. Results, recommendations, justifications, and clear charts

C. Recommendation charts justifications code reviews and results

D. Methodology, code snippets, findings, data tables, and p values

Answer: B

Explanation:

Executive audiences need concise, high-level insights: what you found (results), what you suggest (recommendations), why it matters (justifications), and visual summaries (clear charts). Detailed methods, code, or raw data aren’t appropriate at the C-suite level.

Question: 45

A data scientist is developing a model to predict the outcome of a vote for a national mascot. The choice is between tigers and lions. The full data set represents feedback from individuals representing 17 professions and 12 different locations. The following rank aggregation represents 80% of the data set:

Survey rank	Profession	Location	Voter preference
1	Data scientist	4	Tigers
2	Data scientist	3	Tigers
3	Data analyst	4	Tigers

Which of the following is the most likely concern about the model's ability to predict the outcome of the vote?

A. Interpolated data

B. Extrapolated data

C. In-sample data

D. Out-of-sample data

Answer: D

Explanation:

The aggregated feedback covers only 80% of respondents, mostly from a few professions and locations, so the model hasn't "seen" the remaining 20% (and those underrepresented groups). Its performance on those unseen subsets (out-of-sample data) is therefore the primary concern for how well it will predict the actual vote.

Question: 46

A data scientist is working with a data set that covers a two-year period for a large number of machines. The data set contains:

- Machine system ID numbers
- Sensor measurement values
- Daily time stamps for each machine

The data scientist needs to plot the total measurements from all the machines over the entire time period. Which of the following is the best way to present this data?

A. Scatter plot

B. Line plot

C. Histogram

D. Box-and-whisker plot

Answer: B

Explanation:

Summing measurements across all machines for each day produces a time series, and a line plot is the standard way to visualize how that daily total evolves over the two-year period.

Question: 47

A data scientist has built an image recognition model that distinguishes cars from trucks. The data