EMC E20-007 Exam

Volume: 165 Questions

Question No: 1
You are using MADlib for Linear Regression analysis. Which value does the statement return?
SELECT (linregr(depvar, indepvar)).r2 FROM zeta1;

A. Goodness of fit

B. Coefficients

C. Standard error

D. P-value

Answer: A

Question No: 2
Which data asset is an example of quasi-structured data?

A. Webserver log

B. XML data file

C. Database table

D. News article

Answer: A

Question No: 3
What would be considered "Big Data"?

A. An OLAP Cube containing customer demographic information about 100,000,000 customers

B. Daily Log files from a web server that receives 100,000 hits per minute

C. Aggregated statistical data stored in a relational database table

D. Spreadsheets containing monthly sales data for a Global 100 corporation

Answer: B

Question No: 4
A data scientist plans to classify the sentiment polarity of 10, 000 product reviews collected from the Internet. What is the most appropriate model to use? Suppose labeled training data is available.

A. Na.ve Bayesian classifier

B. Linear regression

C. Logistic regression

D. K-means clustering

Answer: A

Question No: 5
In which lifecycle stage are test and training data sets created?

A. Model building

B. Model planning

C. Discovery

D. Data preparation

Answer: A

Question No: 6
When creating a presentation for a technical audience, what is the main objective?

A. Show that you met the project goals

B. Show how you met the project goals

C. Show if the model will meet the SLA

D. Show the technique to be used in the production environment

Answer: B

Question No: 7
Your company has 3 different sales teams. Each team's sales manager has developed incentive offers to increase the size of each sales transaction. Any sales manager whose incentive program can be shown to increase the size of the average sales transaction will receive a bonus.
Data are available for the number and average sale amount for transactions offering one of the incentives as well as transactions offering no incentive.
The VP of Sales has asked you to determine analytically if any of the incentive programs has resulted in a demonstrable increase in the average sale amount. Which analytical technique would be appropriate in this situation?

A. One-way ANOVA

B. Multi-way ANOVA

C. Student's t-test

D. Wilcoxson Rank Sum Test

Answer: A

Question No: 8
In data visualization, what is used to focus the audience on a key part of a chart?

A. Emphasis colors

B. Detailed text

C. Pastel colors

D. A data table

Answer: A

Question No: 9
Which word or phrase completes the statement? Data-ink ratio is to data visualization as _____ .

A. Confusion matrix is to classifier

B. Data scientist is to big data

C. Seasonality is to ARIMA

D. K-means is to Naive Bayes

Answer: A

Question No: 10
Consider a database with 4 transactions:
Transaction 1: {cheese, bread, milk}
Transaction 2: {soda, bread, milk}
Transaction 3: {cheese, bread}
Transaction 4: {cheese, soda, juice}
You decide to run the association rules algorithm where minimum support is 50%. Which rule has a confidence at least 50%?

A. {cheese} => {bread}

B. {juice} => {cheese}

C. {milk} => {soda}

D. {soda} => {milk}

Answer: A

Question No: 11
You are using the Apriori algorithm to determine the likelihood that a person who owns a home has a good credit score. You have determined that the confidence for the rules used in the algorithm is > 75%. You calculate lift = 1.011 for the rule, "People with good credit are homeowners". What can you determine from the lift calculation?

A. Support for the association is low

B. Leverage of the rules is low

C. The rule is coincidental

D. The rule is true

Answer: C

Question No: 12
Consider a database with 4 transactions:
Transaction 1: {cheese, bread, milk}
Transaction 2: {soda, bread, milk}
Transaction 3: {cheese, bread}
Transaction 4: {cheese, soda, juice}
The minimum support is 25%.
Which rule has a confidence equal to 50%?

A. {bread,milk} => {cheese}

B. {bread} => {milk}

C. {juice} => {soda}

D. {bread} => {cheese}

Answer: D

Question No: 13
Under which circumstance do you need to implement N-fold cross-validation after creating a regression model?

A. There is not enough data to create a test set.

B. The data is unformatted.

C. There are missing values in the data.

D. There are categorical variables in the model.

Answer: A

Question No: 14
What is an appropriate data visualization to use in a presentation for an analyst audience?

A. Pie chart