

Data scientist

 Test duration: 45 minutes

 Difficulty level: **Moderate**



1. What is the primary purpose of exploratory data analysis (EDA)?

- A) To finalize the model to be used
- B) To summarize the main characteristics of the data and uncover patterns, relationships, or anomalies
- C) To deploy the model into production
- D) To perform hypothesis testing

Correct answer: B) To summarize the main characteristics of the data and uncover patterns, relationships, or anomalies

2. Which of the following is a common method for handling missing data in a dataset?

- A) Ignoring the missing data
- B) Deleting the entire dataset
- C) Imputing missing values using the mean, median, or mode
- D) Generating random values for the missing data

Correct answer: C) Imputing missing values using the mean, median, or mode

3. In machine learning, what is overfitting?

- A) A model that performs well on training data but poorly on unseen data due to being too complex
- B) A model that is too simple and underperforms on both training and testing data
- C) A model that only works with large datasets
- D) A model that takes too long to train

Correct answer: A) A model that performs well on training data but poorly on unseen data due to being too complex

4. Which type of machine learning algorithm is used for classification tasks?

- A) Linear Regression
- B) Decision Trees
- C) K-Means Clustering
- D) Principal Component Analysis (PCA)

Correct answer: B) Decision Trees

5. What is the purpose of cross-validation in machine learning?

- A) To test the model on the entire dataset
- B) To ensure that the model generalizes well to unseen data by evaluating it on multiple subsets of the data
- C) To increase the accuracy of the model on the training data
- D) To perform feature selection

Correct answer: B) To ensure that the model generalizes well to unseen data by evaluating it on multiple subsets of the data

6. Which of the following is a common evaluation metric for classification models?

- A) Mean Squared Error (MSE)
- B) R-squared
- C) F1 Score
- D) Adjusted R-squared

Correct answer: C) F1 Score

7. What does the term 'dimensionality reduction' refer to in data science?

- A) Increasing the number of features in the dataset
- B) Reducing the number of features in the dataset to improve model performance and interpretability
- C) Compressing the dataset to save storage space
- D) Aggregating data into fewer categories

Correct answer: B) Reducing the number of features in the dataset to improve model performance and interpretability

8. In natural language processing (NLP), what is the 'Bag of Words' model?

- A) A model that captures the order of words in a sentence
- B) A representation of text data where each document is described by the frequency of words, ignoring grammar and word order
- C) A method for translating languages
- D) A technique for generating random sentences

Correct answer: B) A representation of text data where each document is described by the frequency of words, ignoring grammar and word order

9. What is the purpose of a confusion matrix in evaluating a classification model?

- A) To visualize the relationship between true positive, true negative, false positive, and false negative predictions
- B) To determine the number of features in the dataset
- C) To calculate the model's accuracy
- D) To visualize the distribution of a continuous variable

Correct answer: A) To visualize the relationship between true positive, true negative, false positive, and false negative predictions

10. Which of the following techniques is commonly used for feature selection?

- A) Principal Component Analysis (PCA)
- B) K-Means Clustering
- C) Random Forest Importance
- D) Support Vector Machines (SVM)

Correct answer: C) Random Forest Importance