QUESTION BANK

DATA SCIENCE USING PYTHON

1. What is the primary goal of data preprocessing in data science?
    a) Reducing the dataset size
    b) Enhancing data security
    c) Improving data quality
    d) Increasing computational speed

2. Which library in Python is commonly used for data manipulation and analysis?
    a) Pandas
    b) Numpy
    c) Matplotlib
    d) Scipy

3. What data structure does Pandas introduce for tabular data?
    a) Arrays
    b) DataFrames
    c) Series
    d) Lists

4. Which library is widely used for data visualization in Python?
    a) Numpy
    b) Pandas
    c) Matplotlib
    d) Scipy

5. In Python, which package is used for scientific computing?
    a) Pandas
    b) Numpy
    c) Matplotlib
    d) Seaborn

6. What is the purpose of EDA (Exploratory Data Analysis)?
    a) To create machine learning models
    b) To preprocess data for visualization
    c) To summarize the main characteristics of data
    d) To analyze data in real-time

7. Which library in Python is suitable for creating interactive visualizations?

a) Matplotlib
b) Seaborn
c) Plotly
d) Pandas

8. What is a histogram used for in data visualization?
    a) Displaying time series data
    b) Showing relationships between variables
    c) Representing categorical data
    d) Visualizing data distribution

9. Which statistical measure provides a central value for a dataset?
    a) Mean
    b) Variance
    c) Skewness
    d) Correlation

10. What is a p-value in hypothesis testing?
    a) A measure of data variability
    b) A measure of association between two variables
    c) The probability of observing a test statistic as extreme as the one computed
    d) A measure of data distribution shape

11. Which Python library is commonly used for machine learning tasks?
    a) Matplotlib
    b) Seaborn
    c) Pandas
    d) Scikit-Learn

12. In machine learning, which term refers to the input variables used to predict an outcome?
    a) Target variables
    b) Features
    c) Labels
    d) Observations

13. What is the main difference between supervised and unsupervised learning?
    a) Supervised learning requires labeled data, while unsupervised learning does not.

b) Unsupervised learning requires labeled data, while supervised learning does not.
c) Supervised learning only works with regression tasks.
d) Unsupervised learning only works with classification tasks.

14. What type of machine learning algorithm is used for classifying data into categories?
    a) Regression
    b) Clustering
    c) Dimensionality reduction
    d) Classification

15. Which Python library is often used for deep learning and neural networks?
    a) Numpy
    b) Pandas
    c) Scikit-Learn
    d) TensorFlow

16. Which function is used to split a dataset into training and testing subsets?
    a) train_test_split
    b) split_data
    c) divide_data
    d) test_train_split

17. Which metric is commonly used to evaluate regression models?
    a) F1-score
    b) R-squared
    c) Accuracy
    d) Precision

18. In a confusion matrix, which value represents correctly predicted positive instances?
    a) True Positive (TP)
    b) False Positive (FP)
    c) True Negative (TN)
    d) False Negative (FN)

19. What is the purpose of a decision tree algorithm?
    a) Clustering data points
    b) Finding optimal hyperparameters

c) Identifying outliers in data
d) Making predictions or classifications

20. Which Python library is used for natural language processing (NLP) tasks?
    a) Numpy
    b) Scipy
    c) NLTK
    d) Scikit-Learn

21. Which term refers to reducing the dimensionality of a dataset while retaining important information?
    a) Clustering
    b) Dimensionality reduction
    c) Classification
    d) Regression

22. Which method is used to handle missing data in a dataset?
    a) Removing all missing rows
    b) Replacing missing values with random values
    c) Ignoring missing values during analysis
    d) Imputing missing values with reasonable estimates

23. What is the goal of feature scaling in machine learning?
    a) To reduce the number of features
    b) To standardize features to the same scale
    c) To add more features to the dataset
    d) To remove irrelevant features

24. Which type of machine learning algorithm aims to find patterns in data without using labeled examples?
    a) Supervised learning
    b) Unsupervised learning
    c) Reinforcement learning
    d) Semi-supervised learning

25. Which method is commonly used to validate a machine learning model's performance on unseen data?

a) Testing on the training data
b) Testing on a separate validation dataset
c) Testing on the test dataset
d) Testing on the entire dataset

26. Which Python library provides support for creating interactive dashboards?
    a) Pandas
    b) Matplotlib
    c) Plotly
    d) Seaborn

27. What is the primary purpose of regularization in machine learning?
    a) To overfit the model
    b) To improve training speed
    c) To avoid underfitting by adding complexity
    d) To reduce model complexity and prevent overfitting

28. What is the formula for calculating the mean squared error (MSE) in regression?
    a) $\sqrt{(\Sigma(y - \hat{y})^2 / n)}$
    b) $\Sigma(y - \hat{y})^2 / n$
    c) $1 - \Sigma(y - \hat{y})^2 / n$
    d) $\Sigma(y - \hat{y}) / n$

29. Which algorithm is used for dimensionality reduction while preserving the maximum variance in the data?
    a) Principal Component Analysis (PCA)
    b) k-Nearest Neighbors (k-NN)
    c) Support Vector Machines (SVM)
    d) Linear Regression

30. What is the main difference between supervised and unsupervised machine learning algorithms?
    a) Supervised algorithms require labeled data, while unsupervised algorithms do not.
    b) Unsupervised algorithms require labeled data, while supervised algorithms do not.
    c) Supervised algorithms only work with classification tasks.
    d) Unsupervised algorithms only work with regression tasks.

ANSWERS

1. c) Improving data quality
2. a) Pandas
3. b) DataFrames
4. c) Matplotlib
5. b) Numpy
6. c) Summarizing the main characteristics of data
7. c) Plotly
8. d) Visualizing data distribution
9. a) Mean
10. c) The probability of observing a test statistic as extreme as the one computed
11. d) Scikit-Learn
12. b) Features
13. a) Supervised learning requires labeled data, while unsupervised learning does not.
14. d) Classification
15. d) TensorFlow
16. a) train_test_split
17. b) R-squared
18. a) True Positive (TP)
19. d) Making predictions or classifications
20. c) NLTK
21. b) Dimensionality reduction
22. d) Imputing missing values with reasonable estimates
23. b) To standardize features to the same scale
24. b) Unsupervised learning
25. c) Testing on the test dataset
26. c) Plotly
27. d) To reduce model complexity and prevent overfitting
28. b) $\Sigma(y - \hat{y})^2 / n$
29. a) Principal Component Analysis (PCA)
30. a) Supervised algorithms require labeled data, while unsupervised algorithms do not.

1. What does the term "overfitting" refer to in machine learning?
    a) Model that fits the data accurately
    b) Model that generalizes well to new data
    c) Model that has high bias
    d) Model that performs poorly on training data

2. Which technique is used to evaluate a machine learning model's performance on unseen data during training?
    a) Cross-validation
    b) Testing on training data
    c) Overfitting analysis
    d) Feature selection

3. What is the purpose of the k-fold cross-validation technique?

a) To train the model multiple times
b) To select the optimal learning rate
c) To split the dataset into multiple subsets for testing
d) To prevent overfitting

4. In machine learning, what is the goal of the feature extraction process?
    a) To create new features from existing ones
    b) To reduce the number of features
    c) To select the most important features
    d) To normalize the feature values

5. Which technique is used to handle imbalanced datasets in classification tasks?
    a) Feature scaling
    b) Data imputation
    c) Ensemble learning
    d) Hyperparameter tuning

6. In classification tasks, what does the term "precision" measure?
    a) Ability of the model to correctly identify positive instances
    b) Ability of the model to correctly identify negative instances
    c) Ratio of true positive predictions to total positive predictions
    d) Ratio of true negative predictions to total negative predictions

7. What is the primary purpose of the Naive Bayes algorithm?
    a) Clustering data points
    b) Dimensionality reduction
    c) Solving regression problems
    d) Text classification and spam filtering

8. Which library in Python is commonly used for deep learning and building neural networks?
    a) Numpy
    b) Pandas
    c) Scikit-Learn
    d) TensorFlow

9. In unsupervised learning, what is the primary goal of clustering algorithms?
    a) Predicting numeric values

b) Identifying patterns in labeled data
c) Grouping similar data points into clusters
d) Reducing the dimensionality of the dataset

10. Which type of machine learning algorithm is used for time series forecasting?
    a) Regression
    b) Classification
    c) Clustering
    d) Reinforcement learning

11. What is the primary difference between correlation and causation?
    a) Correlation implies causation.
    b) Causation implies correlation.
    c) Correlation measures a statistical relationship; causation implies a cause-and-effect relationship.
    d) Correlation and causation are synonymous terms.

12. Which Python library provides tools for natural language processing tasks such as tokenization and stemming?
    a) Numpy
    b) Pandas
    c) Scikit-Learn
    d) NLTK

13. What is the primary purpose of the Random Forest algorithm?
    a) Clustering data points
    b) Building decision trees
    c) Solving regression problems
    d) Ensemble learning and improving model performance

14. Which term refers to an optimization technique used to adjust hyperparameters in machine learning models?
    a) Gradient Descent
    b) Feature Scaling
    c) Cross-validation
    d) Grid Search

15. Which technique is used to handle multicollinearity in regression analysis?
    a) Ensemble learning
    b) Principal Component Analysis (PCA)
    c) Cross-validation
    d) Regularization

16. What is the purpose of the elbow method in clustering?
    a) It helps determine the number of clusters in the dataset.
    b) It identifies outliers in the data.
    c) It calculates the silhouette score of the clusters.
    d) It measures the accuracy of the clustering algorithm.

17. In time series analysis, what does the term "lag" refer to?
    a) The time interval between data points
    b) The number of data points in the dataset
    c) The difference between predicted and actual values
    d) The number of time steps in the past to consider for analysis

18. Which Python library provides tools for creating interactive visualizations with declarative syntax?
    a) Matplotlib
    b) Seaborn
    c) Plotly
    d) Pandas

19. What is the primary purpose of the Chi-Square test?
    a) Testing the equality of means in two samples
    b) Testing the relationship between categorical variables
    c) Calculating the correlation coefficient
    d) Evaluating the fit of a regression model

20. Which technique is used to mitigate the curse of dimensionality in high-dimensional datasets?
    a) Ensemble learning
    b) Feature scaling
    c) Dimensionality reduction
    d) Hyperparameter tuning

ANSWERS

1. a) Model that fits the data accurately
2. a) Cross-validation
3. c) To split the dataset into multiple subsets for testing
4. a) To create new features from existing ones
5. c) Ensemble learning
6. a) Ability of the model to correctly identify positive instances
7. d) Text classification and spam filtering
8. d) TensorFlow
9. c) Grouping similar data points into clusters
10. a) Regression
11. c) Correlation measures a statistical relationship; causation implies a cause-and-effect relationship.
12. d) NLTK
13. d) Ensemble learning and improving model performance
14. d) Grid Search
15. d) Regularization
16. a) It helps determine the number of clusters in the dataset.
17. d) The number of time steps in the past to consider for analysis
18. c) Plotly
19. b) Testing the relationship between categorical variables
20. c) Dimensionality reduction

1. Which of the following are common measures of central tendency?
    a) Mean
    b) Median
    c) Mode
    d) Variance

2. Which of the following Python libraries can be used for dimensionality reduction?
    a) Scikit-Learn
    b) Numpy
    c) TensorFlow
    d) PCA

3. Which of the following are steps in the data preprocessing phase?
    a) Exploratory Data Analysis (EDA)
    b) Data cleaning and transformation
    c) Hyperparameter tuning
    d) Model selection

4. Which of the following algorithms are used for unsupervised learning?
    a) Decision Trees
    b) K-Means Clustering
    c) Linear Regression
    d) Support Vector Machines (SVM)

5. Which of the following methods can be used for feature selection?
    a) Recursive Feature Elimination (RFE)
    b) Principal Component Analysis (PCA)
    c) LASSO (Least Absolute Shrinkage and Selection Operator)
    d) Ridge Regression

6. Which of the following Python libraries are commonly used for time series analysis?
    a) Numpy
    b) Pandas
    c) Statsmodels
    d) Matplotlib

7. Which of the following evaluation metrics can be used for regression models?
    a) Mean Absolute Error (MAE)
    b) F1-score
    c) Precision
    d) R-squared

8. Which of the following Python libraries are used for natural language processing (NLP) tasks?
    a) NLTK
    b) Scikit-Learn
    c) SpaCy
    d) Plotly

9. Which of the following statements about the k-Nearest Neighbors (k-NN) algorithm are true?
    a) It's a supervised learning algorithm.
    b) It's used for clustering data.
    c) It makes predictions based on the majority class of its k-nearest neighbors.
    d) It's sensitive to the scale of features.

10. Which of the following techniques can help address the issue of multicollinearity in regression analysis?
    a) Ridge Regression
    b) Decision Trees
    c) K-Means Clustering
    d) Logistic Regression

ANSWERS

1. a) Mean, b) Median, c) Mode
2. a) Scikit-Learn, d) PCA
3. a) Exploratory Data Analysis (EDA), b) Data cleaning and transformation
4. b) K-Means Clustering
5. a) Recursive Feature Elimination (RFE), b) Principal Component Analysis (PCA), c) LASSO (Least Absolute Shrinkage and Selection Operator)
6. b) Pandas, c) Statsmodels
7. a) Mean Absolute Error (MAE), d) R-squared
8. a) NLTK, c) SpaCy
9. c) It makes predictions based on the majority class of its k-nearest neighbors, d) It's sensitive to the scale of features.
10. a) Ridge Regression