**Exploratory Data Analysis (EDA)**:

**Question 1:** What is the primary goal of Exploratory Data Analysis (EDA)?

A) To build predictive models
B) To summarize and visualize the key features of a dataset
C) To check the accuracy of a dataset
D) To identify outliers only

**Answer:** B) To summarize and visualize the key features of a dataset

**Question 2:** Which of the following is not a typical visualization used in EDA?

A) Histogram
B) Box plot
C) Decision tree
D) Scatter plot

**Answer:** C) Decision tree

**Question 3:** What is the purpose of a box plot in EDA?

A) To show the correlation between variables
B) To visualize the distribution and detect outliers
C) To identify trends in the data
D) To calculate summary statistics

**Answer:** B) To visualize the distribution and detect outliers

**Question 4:** Which statistical measure is commonly used in EDA to summarize the central tendency of a numerical variable?

A) Mode
B) Median
C) Mean
D) All of the above

**Answer:** D) All of the above

**Question 5:** What does a correlation matrix show during EDA?

A) The distribution of each variable
B) The relationship between pairs of variables
C) The summary statistics of the dataset
D) The outliers in the data

**Answer:** B) The relationship between pairs of variables

**Question 6:** Which technique is commonly used for handling missing data in EDA?

A) Linear regression
B) Imputation
C) K-means clustering
D) Hypothesis testing

**Answer:** B) Imputation

**Question 7:** In a scatter plot, what does a strong linear relationship between two variables indicate?

A) No relationship between the variables
B) A positive or negative correlation
C) The variables are categorical
D) A weak correlation

**Answer:** B) A positive or negative correlation

**Question 8:** When performing EDA, which of the following is most useful for detecting skewness in a numerical variable?

A) Histogram
B) Line plot
C) Heatmap
D) Pie chart

**Answer:** A) Histogram

**Question 9:** Which of the following methods can be used to detect outliers in a dataset?

A) Scatter plot
B) Box plot
C) Histogram
D) All of the above

**Answer:** D) All of the above

**Question 10:** Which type of EDA plot would be most appropriate to visualize the relationship between a categorical variable and a numerical variable?

A) Histogram

B) Box plot
C) Heatmap
D) Bar chart

**Answer:** B) Box plot

**Multiple Linear Regression**

**Question 1:** In a multiple linear regression model, what does the equation $y=\beta0+\beta1x1+\beta2x2+...+\betanxn+\epsilon$ y = \beta_0 + \beta_1x_1 + \beta_2x_2 + ... + \beta_nx_n + \epsilon represent?

A) The relationship between one dependent variable and multiple independent variables
B) The relationship between multiple dependent variables and independent variables
C) The correlation between variables
D) The prediction of categorical values

**Answer:** A) The relationship between one dependent variable and multiple independent variables

**Question 2:** Which of the following libraries is commonly used for implementing multiple linear regression in Python?

A) TensorFlow
B) scikit-learn
C) NumPy
D) Keras

**Answer:** B) scikit-learn

**Question 3:** What is the purpose of the fit() function in scikit-learn's LinearRegression?

A) To evaluate the performance of the model
B) To split the dataset into training and testing sets
C) To train the model on the data
D) To visualize the regression line

**Answer:** C) To train the model on the data

**Question 4:** In multiple linear regression, what assumption must be true for the model to be valid?

A) The features are categorical
B) The data is homoscedastic

C) The dependent variable is binary
D) The independent variables must be independent

**Answer:** B) The data is homoscedastic

**Question 5:** Which of the following metrics is commonly used to evaluate the performance of a multiple linear regression model?

A) Accuracy
B) R-squared
C) AUC
D) F1-score

**Answer:** B) R-squared

**Question 6:** What does a high R-squared value in multiple linear regression indicate?

A) The model has a high bias
B) The model does not explain much of the variance in the dependent variable
C) The model explains a large proportion of the variance in the dependent variable
D) The model is overfitting the data

**Answer:** C) The model explains a large proportion of the variance in the dependent variable

**Question 7:** In the context of multiple linear regression, what does multicollinearity refer to?

A) When the model is overfitting the training data
B) When two or more independent variables are highly correlated
C) When the dependent variable is categorical
D) When the data is not normally distributed

**Answer:** B) When two or more independent variables are highly correlated

**Binary Logistic Regression**

**Question 8:** What type of variable does binary logistic regression predict?

A) Continuous numerical variables
B) Binary categorical variables (0 or 1)
C) Multiclass categorical variables
D) Both binary and continuous variables

**Answer:** B) Binary categorical variables (0 or 1)

**Question 9:** In binary logistic regression, what is the output of the sigmoid function?

A) A probability score between 0 and 1
B) A continuous value between 0 and infinity
C) A binary value (0 or 1)
D) A regression coefficient

**Answer:** A) A probability score between 0 and 1

**Question 10:** Which of the following functions in scikit-learn is used for binary logistic regression?

A) LinearRegression()
B) LogisticRegression()
C) SVC()
D) KMeans()

**Answer:** B) LogisticRegression()

**Question 11:** Which of the following is a key assumption of logistic regression?

A) The dependent variable is normally distributed
B) The relationship between the dependent and independent variables is linear
C) The dependent variable is binary
D) The residuals are normally distributed

**Answer:** C) The dependent variable is binary

**Question 12:** In the context of binary logistic regression, what does a logistic function (sigmoid curve) help to do?

A) Minimize the loss function
B) Transform the output to a probability
C) Calculate the decision boundary
D) Normalize the data

**Answer:** B) Transform the output to a probability

**Question 13:** What is the main purpose of using the predict() method in a logistic regression model?

A) To train the model
B) To evaluate the performance of the model
C) To make predictions on unseen data
D) To calculate the coefficients of the model

**Answer:** C) To make predictions on unseen data

**Decision Trees**

**Question 14:** Which of the following is a key advantage of using decision trees for classification?

A) High bias and low variance
B) They are easy to interpret and visualize
C) They require data normalization
D) They cannot handle categorical variables

**Answer:** B) They are easy to interpret and visualize

**Question 15:** Which criterion is commonly used to split nodes in a decision tree for classification tasks?

A) Mean squared error
B) Entropy or Gini impurity
C) R-squared
D) Log-likelihood

**Answer:** B) Entropy or Gini impurity

**Question 16:** In decision trees, which of the following would be considered a "leaf node"?

A) A node that represents a decision or split in the data
B) A node where the data is not further split and a prediction is made
C) A node used to calculate the Gini index
D) A node where the data is split into more subsets

**Answer:** B) A node where the data is not further split and a prediction is made

**Question 17:** Which Python library is commonly used for implementing decision trees?

A) scikit-learn
B) TensorFlow
C) Matplotlib
D) Keras

**Answer:** A) scikit-learn

**Question 18:** What is the purpose of pruning a decision tree?

A) To add more splits to the tree
B) To remove nodes that provide little predictive value and reduce overfitting
C) To increase the depth of the tree
D) To decrease the size of the dataset

**Answer:** B) To remove nodes that provide little predictive value and reduce overfitting

**Question 19:** Which of the following is the primary disadvantage of using decision trees?

A) They are highly interpretable
B) They are prone to overfitting, especially with deep trees
C) They perform poorly on unstructured data
D) They cannot handle both numerical and categorical variables

**Answer:** B) They are prone to overfitting, especially with deep trees

**Question 20:** What is the max_depth parameter used for in scikit-learn's DecisionTreeClassifier?

A) To specify the maximum number of leaf nodes
B) To define the maximum depth of the tree
C) To determine the minimum samples required to split a node
D) To define the number of features to consider when splitting a node

**Answer:** B) To define the maximum depth of the tree

**Natural Language Processing (NLP) in Python**:

**Question 1:**

Which Python library is commonly used for Natural Language Processing (NLP)?

A) scikit-learn
B) NLTK (Natural Language Toolkit)
C) TensorFlow
D) PyTorch

**Answer:** B) NLTK (Natural Language Toolkit)

**Question 2:**

What is the primary function of the word_tokenize() method in NLTK?

A) To remove stop words
B) To split text into words
C) To stem words
D) To convert text to lowercase

**Answer:** B) To split text into words

**Question 3:**

Which of the following is used for stemming words in NLTK?

A) WordNetLemmatizer()
B) PorterStemmer()
C) stopwords()
D) countvectorizer()

**Answer:** B) PorterStemmer()

**Question 4:**

What is the difference between stemming and lemmatization in NLP?

A) Stemming reduces words to their root forms, while lemmatization returns words to their dictionary form
B) Stemming and lemmatization are the same
C) Stemming uses a lexicon, while lemmatization uses rules
D) Lemmatization is faster than stemming

**Answer:** A) Stemming reduces words to their root forms, while lemmatization returns words to their dictionary form

**Question 5:**

Which method is used to remove stop words from a text in NLTK?

A) stopwords.remove()
B) nltk.corpus.stopwords.words()
C) word_tokenize()
D) wordnet.synsets()

**Answer:** B) nltk.corpus.stopwords.words()

**Question 6:**

In Python's spaCy library, which object is used to represent the entire

processed document, including tokens, sentences, and entities?

A) Doc
B) Token
C) SpacyText
D) TextBlob

**Answer:** A) Doc

**Question 7:**

Which of the following methods in the CountVectorizer class from scikit-learn is used to convert a collection of text documents into a matrix of token counts?

A) fit()
B) transform()
C) fit_transform()
D) vectorize()

**Answer:** C) fit_transform()

**Question 8:**

What does Named Entity Recognition (NER) in NLP aim to identify?

A) The grammatical structure of sentences
B) The sentiment of a text
C) The named entities such as people, locations, and organizations
D) The syntactic role of words in a sentence

**Answer:** C) The named entities such as people, locations, and organizations

**Question 9:**

Which Python library is commonly used to perform part-of-speech (POS) tagging?

A) pandas
B) NLTK
C) NumPy
D) Matplotlib

**Answer:** B) NLTK

**Question 10:**

Which of the following methods in the TextBlob library is used for sentiment

analysis?

A) sentiment.polarity()
B) analyze_sentiment()
C) polarity_score()
D) tokenize()

**Answer:** A) sentiment.polarity()

**Image Processing with Convolutional Neural Networks (CNNs)**:

**Question 1:**

What is the primary purpose of a convolutional layer in a CNN model?

A) To reduce the dimensions of the input
B) To apply filters for feature extraction
C) To perform classification of input data
D) To combine multiple models

**Answer:** B) To apply filters for feature extraction

**Question 2:**

What is the role of pooling layers in CNNs?

A) To add more parameters to the model
B) To extract features using convolution
C) To reduce the spatial dimensions of feature maps
D) To increase the resolution of the input

**Answer:** C) To reduce the spatial dimensions of feature maps

**Question 3:**

Which of the following operations is commonly used in a max-pooling layer?

A) Computing the sum of all values in a region
B) Taking the average of all values in a region
C) Selecting the maximum value in a region
D) Subtracting the minimum value from the maximum

**Answer:** C) Selecting the maximum value in a region

**Question 4:**

In the context of CNNs, what does the term "stride" refer to?

A) The size of the filter applied during convolution
B) The step size with which the filter moves across the input
C) The depth of the output feature maps
D) The number of parameters in the network

**Answer:** B) The step size with which the filter moves across the input

**Question 5:**

What is the ReLU activation function used for in CNNs?

A) To normalize the input data
B) To introduce non-linearity by setting negative values to zero
C) To reduce the dimensionality of feature maps
D) To optimize the weights during training

**Answer:** B) To introduce non-linearity by setting negative values to zero

**Question 6:**

What is the purpose of the fully connected layer in a CNN?

A) To perform spatial feature extraction
B) To combine all extracted features for final prediction
C) To reduce the size of feature maps
D) To calculate the convolution operation

**Answer:** B) To combine all extracted features for final prediction

**Question 7:**

Which technique is commonly used to prevent overfitting in CNN models?

A) Increasing the filter size
B) Applying dropout regularization
C) Reducing the number of layers
D) Removing the pooling layers

**Answer:** B) Applying dropout regularization

**Question 8:**

What is the purpose of using padding in a CNN?

A) To increase the size of the filters
B) To adjust the input size for pooling layers

C) To prevent dimensionality reduction during convolution
D) To normalize feature maps

**Answer:** C) To prevent dimensionality reduction during convolution

**Question 9:**

Which of the following optimizers is commonly used in CNN models for training?

A) Stochastic Gradient Descent (SGD)
B) K-means Clustering
C) Support Vector Machine
D) Principal Component Analysis

**Answer:** A) Stochastic Gradient Descent (SGD)

**Question 10:**

What is the primary role of a softmax layer in a CNN model?

A) To extract features from images
B) To reduce overfitting
C) To convert logits into probabilities for classification
D) To compute the loss function

**Answer:** C) To convert logits into probabilities for classification