Multiple-Choice Questions (MCQs) in English:

- **Which of the following is a characteristic of Big Data?**
  - A) Simplicity
  - B) Velocity
  - C) Predictability
  - D) Low cost

  Answer: B) Velocity

- **What is NOT a component of Data Science?**
  - A) Machine Learning
  - B) Data Analysis
  - C) Field Surveys
  - D) Big Data Processing

  Answer: C) Field Surveys

- **What is the purpose of Feature Engineering?**
  - A) To collect data
  - B) To clean the data
  - C) To improve model performance
  - D) To visualize data

  Answer: C) To improve model performance

- **Which of the following tools is commonly used for Exploratory Data Analysis (EDA)?**
  - A) Microsoft Word
  - B) Python's Seaborn
  - C) AutoCAD
  - D) MongoDB

  Answer: B) Python's Seaborn

- **What technique is used to encode categorical variables into numerical ones?**
  - A) Principal Component Analysis
  - B) Clustering
  - C) One-Hot Encoding
  - D) Normalization

  Answer: C) One-Hot Encoding

- **What is the primary goal of the CRISP-DM methodology?**
  - A) To manage data storage efficiently
  - B) To provide a structured process for data science projects
  - C) To automate machine learning workflows
  - D) To collect data from IoT sensors

  Answer: B) To provide a structured process for data science projects

- **Which of the following is NOT a phase in the CRISP-DM process?**
  - A) Business Understanding
  - B) Deployment
  - C) Data Archiving
  - D) Modeling

  Answer: C) Data Archiving

- **Which method is commonly used to detect outliers in data?**
  - A) Histogram
  - B) Z-Score
  - C) Scatter Plot
  - D) Pivot Table

  Answer: B) Z-Score

- **Why is data preprocessing essential?**
  - A) It eliminates the need for modeling
  - B) It improves data quality and consistency
  - C) It increases the volume of data
  - D) It guarantees 100% accuracy in predictions

  Answer: B) It improves data quality and consistency

- **What does One-Hot Encoding do?**
  - A) Reduces the size of the dataset
  - B) Converts categorical variables into binary features
  - C) Normalizes numerical variables
  - D) Handles missing values

  Answer: B) Converts categorical variables into binary features

- **Which of the following is an example of supervised learning?**
  - A) Clustering
  - B) Linear Regression
  - C) Principal Component Analysis (PCA)
  - D) Data Transformation

    Answer: B) Linear Regression

- **What is a common method for handling missing data?**
  - A) Removing all rows
  - B) Replacing with mean or median values
  - C) Ignoring the issue
  - D) Adding random values

    Answer: B) Replacing with mean or median values

13. **What does the term "Exploratory Data Analysis (EDA)" focus on?**

- A) Automating data collection
- B) Identifying patterns and relationships in data
- C) Building machine learning models
- D) Writing SQL queries

  Answer: B) Identifying patterns and relationships in data

- **Which visualization technique is best for identifying the relationship between two variables?**
  - A) Histogram
  - B) Scatter Plot
  - C) Box Plot
  - D) Pie Chart

    Answer: B) Scatter Plot

- **What is the main advantage of Normalization in data preprocessing?**
  - A) It eliminates missing data completely
  - B) It ensures all features have the same range
  - C) It simplifies categorical variables
  - D) It removes outliers

    Answer: B) It ensures all features have the same range

- **What does Feature Engineering primarily involve?**
  - A) Collecting raw data
  - B) Cleaning and transforming features to improve model performance
  - C) Deploying models to production
  - D) Selecting algorithms

    Answer: B) Cleaning and transforming features to improve model performance

- **Which of the following is a method to reduce overfitting in models?**
  - A) Using a smaller dataset
  - B) Leave-one-out encoding
  - C) Ignoring outliers
  - D) Adding more features

    Answer: B) Leave-one-out encoding

- **What is a "Target Variable" in a dataset?**
  - A) A variable used to encode categories
  - B) The dependent variable being predicted
  - C) An independent variable in regression analysis
  - D) A feature used for clustering

    Answer: B) The dependent variable being predicted

- **Which of the following describes PCA (Principal Component Analysis)?**
  - A) A method for data cleaning
  - B) A dimensionality reduction technique
  - C) A clustering algorithm
  - D) A data encoding method

    Answer: B) A dimensionality reduction technique

- **What is the purpose of Cross-Validation in machine learning?**
  - A) To handle missing values
  - B) To evaluate model performance on unseen data
  - C) To reduce dataset size
  - D) To select the best feature encoding technique

    Answer: B) To evaluate model performance on unseen data

- **Which encoding technique is best when categorical variables have a high cardinality?**
  - A) One-Hot Encoding
  - B) Frequency Encoding
  - C) Standardization
  - D) Log Transformation

  Answer: B) Frequency Encoding

- **Why is Feature Scaling important for some machine learning algorithms?**
  - A) It reduces noise in the dataset
  - B) It ensures features have similar influence on the model
  - C) It eliminates categorical variables
  - D) It speeds up model deployment

  Answer: B) It ensures features have similar influence on the model

- **Which of the following is a characteristic of unsupervised learning?**
  - A) It uses labeled data for training
  - B) It aims to cluster or group data without labels
  - C) It builds predictive models
  - D) It requires a target variable

  Answer: B) It aims to cluster or group data without labels

- **What is the significance of the "3 Vs" in Big Data?**
  - A) They describe the types of data
  - B) They define the characteristics of Big Data (Volume, Velocity, Variety)
  - C) They represent the stages of data preprocessing
  - D) They are metrics for model evaluation

  Answer: B) They define the characteristics of Big Data (Volume, Velocity, Variety)

- **Which technique is commonly used to visualize correlations between multiple variables?**
  - A) Pie Chart
  - B) Correlation Heatmap
  - C) Histogram
  - D) Line Graph

  Answer: B) Correlation Heatmap

- **What is the primary purpose of the MQTT protocol in IoT data collection?**
  a) Reducing energy consumption
  b) Transmitting data to a central broker
  c) Creating synthetic samples
  d) Improving model accuracy

  Answer: b) Transmitting data to a central broker

- **Which of the following is an example of data rebalancing through oversampling?**
  a) Removing samples from the majority class
  b) Adding new devices to collect data
  c) Using SMOTE to generate synthetic examples
  d) Modifying class weights in the algorithm

  Answer: c) Using SMOTE to generate synthetic examples

- **What type of data source is directly collected from surveys or experiments?**
  a) Secondary data
  b) Primary data
  c) IoT data
  d) External APIs

  Answer: b) Primary data

- **What does the Arduino Uno board include to support microcontroller functionality?**
  a) Built-in Wi-Fi module
  b) Flash memory, SRAM, and EEPROM
  c) Pre-installed sensors
  d) Machine learning algorithms

  Answer: b) Flash memory, SRAM, and EEPROM

- **Which sampling technique aims to reduce the dataset size by removing samples from the majority class?**
  a) SMOTE
  b) Random undersampling
  c) Weighted sampling
  d) API-based sampling

  Answer: b) Random undersampling

- **What does the term "feature selection" in data science refer to?**
  a) Collecting data from IoT sensors
  b) Choosing the most relevant variables for the model
  c) Adjusting model thresholds
  d) Balancing the dataset using synthetic techniques

  Answer: b) Choosing the most relevant variables for the model

- **What is a key characteristic of IoT data?**
  a) It is primarily analog.
  b) It is collected in real-time.
  c) It requires manual processing.
  d) It cannot be used for predictive modeling.

  Answer: b) It is collected in real-time.

- **Which metric measures the ratio of correctly predicted positive instances to the total actual positives?**
  a) Precision
  b) Recall (Sensitivity)
  c) Specificity
  d) F1-Score

  Answer: b) Recall (Sensitivity)

- **What is a common drawback of imbalanced datasets in machine learning?**
  a) They always require feature selection.
  b) Models tend to ignore the minority class.
  c) They are unsuitable for binary classification.
  d) They have fewer features to train the model.

Answer: b) Models tend to ignore the minority class.

- **What does SMOTE stand for?**
  a) Simple Minority Oversampling Technique
  b) Synthetic Minority Oversampling Technique
  c) Systematic Model Optimization Technique
  d) Structured Model Output Training Experiment

  Answer: b) Synthetic Minority Oversampling Technique

- **Which approach involves modifying the model to reduce misclassification costs?**
  a) Cost-sensitive learning
  b) Random oversampling
  c) API-based data collection
  d) Threshold adjustment

  Answer: a) Cost-sensitive learning

- **Which of the following is an example of a connectivity protocol for IoT devices?**
  a) Python
  b) LoRaWAN
  c) SMOTE
  d) EEPROM

  Answer: b) LoRaWAN

- **How can logistic regression be adapted to handle imbalanced datasets?**
  a) Using class weights in the algorithm
  b) Reducing the threshold to 0.25
  c) Increasing the number of training epochs
  d) Applying SMOTE to the test data

  Answer: a) Using class weights in the algorithm

- **What does the precision metric evaluate?**
  a) The proportion of actual positives correctly identified
  b) The proportion of predicted positives that are correct
  c) The total number of samples classified correctly
  d) The sensitivity of the classification threshold

  Answer: b) The proportion of predicted positives that are correct

- **What type of IoT device is typically used to measure soil moisture?**
  a) Actuator
  b) Capacitive sensor
  c) Motion detector
  d) Temperature sensor

  Answer: b) Capacitive sensor
