Data science mcq

1. **What is the purpose of the pandas library in Python?**
    - A) Data visualization
    - B) Web development
    - C) Data manipulation and analysis
    - D) Machine learning
    - **Answer: C**
2. **Which of the following is a supervised learning algorithm?**
    - A) K-means clustering
    - B) Principal Component Analysis (PCA)
    - C) Linear Regression
    - D) t-Distributed Stochastic Neighbor Embedding (t-SNE)
    - **Answer: C**
3. **In a classification problem, what is the purpose of the confusion matrix?**
    - A) To determine the most accurate model
    - B) To evaluate the performance of a classification model
    - C) To visualize data distribution
    - D) To handle missing values
    - **Answer: B**
4. **Which of the following metrics is used to evaluate the performance of a regression model?**
    - A) Accuracy
    - B) F1 Score
    - C) Mean Absolute Error (MAE)
    - D) Precision
    - **Answer: C**
5. **What does 'overfitting' mean in the context of machine learning?**
    - A) The model performs well on the training data but poorly on new data.
    - B) The model performs well on both training and new data.
    - C) The model is too simple to capture the underlying pattern.
    - D) The model requires more features.
    - **Answer: A**

6. **Which algorithm is used for dimensionality reduction?**
    - A) Random Forest
    - B) Logistic Regression
    - C) Principal Component Analysis (PCA)
    - D) Decision Tree
    - **Answer: C**
7. **What is the purpose of the train_test_split function in scikit-learn?**
    - A) To split data into training and testing sets
    - B) To normalize data
    - C) To select features
    - D) To perform cross-validation
    - **Answer: A**
8. **In a decision tree, what is a 'leaf node'?**
    - A) A node that splits the data
    - B) A node representing a decision
    - C) A node representing a class label or output
    - D) A node with multiple branches
    - **Answer: C**
9. **What is 'feature scaling'?**
    - A) Increasing the number of features in the dataset
    - B) Normalizing or standardizing features so they contribute equally to the model
    - C) Removing irrelevant features
    - D) Combining features into a single feature
    - **Answer: B**
10. **What does 'cross-validation' help with?**
    - A) Reducing bias
    - B) Increasing the dataset size
    - C) Selecting the best features
    - D) Assessing how the model generalizes to an independent dataset
    - **Answer: D**
11. **Which of the following is an unsupervised learning algorithm?**

- A) K-Nearest Neighbors
- B) Support Vector Machine
- C) K-means Clustering
- D) Decision Tree
- **Answer: C**

12. **What is 'regularization' in machine learning?**
    - A) A technique to reduce the complexity of a model to prevent overfitting
    - B) A method to handle missing values
    - C) A way to increase the model's accuracy
    - D) A process to normalize data
    - **Answer: A**
13. **What does 'bagging' stand for in ensemble methods?**
    - A) Boosted Aggregation
    - B) Bootstrap Aggregating
    - C) Balanced Aggregating
    - D) Binary Aggregating
    - **Answer: B**
14. **What is a ROC curve used for?**
    - A) To visualize the performance of a classification model
    - B) To perform regression analysis
    - C) To check the variance in a dataset
    - D) To handle missing values
    - **Answer: A**
15. **Which metric is used to evaluate the performance of a classification model by comparing predicted and actual values?**
    - A) Mean Squared Error (MSE)
    - B) Root Mean Squared Error (RMSE)
    - C) Precision
    - D) Recall
    - **Answer: C**
16. **What is 'ensemble learning'?**

- A) Combining predictions from multiple models to improve accuracy
- B) Training a single model with multiple datasets
- C) Using a single algorithm for multiple tasks
- D) Selecting features based on their importance
- **Answer: A**

17. **Which of the following is NOT a type of cross-validation?**
    - A) K-Fold Cross-Validation
    - B) Leave-One-Out Cross-Validation
    - C) Stratified Cross-Validation
    - D) Gradient Cross-Validation
    - **Answer: D**
18. **What is the main goal of Principal Component Analysis (PCA)?**
    - A) To classify data into categories
    - B) To reduce the dimensionality of data while retaining variance
    - C) To predict future values
    - D) To cluster similar data points
    - **Answer: B**
19. **What does 'bootstrap sampling' involve?**
    - A) Sampling with replacement to create multiple datasets
    - B) Removing samples to create a smaller dataset
    - C) Combining samples from different datasets
    - D) Normalizing samples before training
    - **Answer: A**
20. **Which metric would you use to measure how many times a positive class was correctly predicted?**
    - A) True Positive Rate
    - B) False Positive Rate
    - C) False Negative Rate
    - D) True Negative Rate
    - **Answer: A**
21. **What is a 'hyperparameter' in machine learning?**

- A) A parameter that is learned during the training process
- B) A parameter set before training begins and controls the learning process
- C) A parameter that measures model accuracy
- D) A parameter used to normalize data
- **Answer: B**

22. **Which method is used to handle missing values in a dataset?**
    - A) Imputation
    - B) Normalization
    - C) Standardization
    - D) Feature Selection
    - **Answer: A**
23. **In a neural network, what is the purpose of an activation function?**
    - A) To scale input features
    - B) To introduce non-linearity into the model
    - C) To reduce the number of features
    - D) To calculate the loss
    - **Answer: B**
24. **Which of the following is a type of regression algorithm?**
    - A) Support Vector Machines (SVM)
    - B) K-Nearest Neighbors (KNN)
    - C) Ridge Regression
    - D) K-means Clustering
    - **Answer: C**

25. **What is 'gradient descent'?**
    - A) An optimization algorithm to minimize the cost function
    - B) A method to standardize features
    - C) A technique for dimensionality reduction
    - D) An algorithm for classification
    - **Answer: A**
26. **What does the 'F1 Score' combine?**
    - A) Precision and Recall

- B) Accuracy and Precision
- C) Recall and Specificity
- D) Precision and Specificity
- **Answer: A**

27. **Which type of machine learning is used when the outcome variable is categorical?**
    - A) Supervised Learning
    - B) Unsupervised Learning
    - C) Reinforcement Learning
    - D) Semi-Supervised Learning
    - **Answer: A**
28. **What is 'data augmentation'?**
    - A) Increasing the amount of data by creating variations of existing data
    - B) Reducing the size of the dataset
    - C) Combining multiple datasets into one
    - D) Standardizing the dataset
    - **Answer: A**
29. **In a time series analysis, what is 'seasonality'?**
    - A) The trend observed over long periods
    - B) Regular patterns or cycles within a fixed period
    - C) Random noise in the data
    - D) The trend observed in short periods
    - **Answer: B**
30. **What is 'Data Wrangling'?**
    - A) Cleaning and transforming raw data into a usable format
    - B) Visualizing data
    - C) Performing statistical analysis
    - D) Training a machine learning model
    - **Answer: A**
31. **What is the purpose of a 'confusion matrix'?**
    - A) To evaluate the performance of a classification algorithm
    - B) To visualize data distributions

- C) To scale numerical features
- D) To reduce data dimensionality
- **Answer: A**

32. **Which of the following techniques is used for feature selection?**
    - A) Recursive Feature Elimination (RFE)
    - B) Gradient Descent
    - C) Cross-Validation
    - D) Principal Component Analysis (PCA)
    - **Answer: A**
33. **In clustering, what does 'silhouette score' measure?**
    - A) The goodness of fit for the clusters
    - B) The distance between different clusters
    - C) The density of clusters
    - D) The separation distance between clusters
    - **Answer: D**
34. **Which of the following is used for anomaly detection?**
    - A) Isolation Forest
    - B) K-Means Clustering
    - C) Principal Component Analysis (PCA)
    - D) Linear Regression
    - **Answer: A**
35. **What is 'dimensionality reduction'?**
    - A) Reducing the number of features in a dataset while retaining important information
    - B) Increasing the size of the dataset
    - C) Combining features into a new feature
    - D) Removing noisy features from the dataset
    - **Answer: A**
36. **What does 'L1 regularization' help with?**
    - A) Shrinkage of feature coefficients
    - B) Adding more features

- C) Reducing the learning rate
- D) Reducing the number of clusters
- **Answer: A**

37. **Which metric is used to evaluate the performance of a regression model?**
    - A) Mean Absolute Error (MAE)
    - B) Precision
    - C) Recall
    - D) Accuracy
    - **Answer: A**
38. **What is 'ensemble learning'?**
    - A) Combining multiple models to improve performance
    - B) Training a single model with multiple datasets
    - C) Using different algorithms for different tasks
    - D) Combining features to create a new feature
    - **Answer: A**
39. **What does the 'R-squared' value represent in a regression model?**
    - A) The proportion of variance explained by the model
    - B) The accuracy of the model
    - C) The error rate of the model
    - D) The precision of the model
    - **Answer: A**
40. **In a neural network, what is the purpose of dropout?**
    - A) To prevent overfitting by randomly dropping neurons during training
    - B) To increase the learning rate
    - C) To normalize input data
    - D) To combine multiple neural networks
    - **Answer: A**
41. **What is a 'ROC curve'?**
    - A) A plot of the true positive rate against the false positive rate
    - B) A plot of accuracy against the number of features
    - C) A plot of loss against the number of iterations

- D) A plot of precision against recall
- **Answer: A**

42. **Which of the following is a type of unsupervised learning?**
    - A) K-means Clustering
    - B) Logistic Regression
    - C) Decision Trees
    - D) Random Forest
    - **Answer: A**
43. **What does 'cross-validation' assess?**

    * 'cross-validation' means generalization

    - A) The generalization ability of a model
    - B) The training speed of a model
    - C) The feature importance
    - D) The model complexity
    - **Answer: A**
44. **What does 'Hyperparameter tuning' involve?**
    - A) Adjusting parameters set before training to improve model performance
    - B) Optimizing the training algorithm
    - C) Reducing the number of features
    - D) Normalizing the data
    - **Answer: A**
45. **Which technique is commonly used for data visualization?**
    - A) PCA
    - B) t-SNE
    - C) K-means
    - D) Random Forest
    - **Answer: B**

46. **What is the purpose of 'feature engineering'?**
    - A) Creating new features or modifying existing ones to improve model performance
    - B) Selecting a subset of features for training
    - C) Normalizing the feature values
    - D) Removing irrelevant features

- **Answer: A**

47. **What is 'Label Encoding'?**
    - A) Converting categorical labels into numerical format
    - B) Normalizing feature values
    - C) Reducing the number of features
    - D) Encoding text data into feature vectors
    - **Answer: A**
48. **What is 'One-Hot Encoding'?**
    - A) Converting categorical variables into binary vectors
    - B) Normalizing numerical features
    - C) Reducing the number of categories
    - D) Scaling feature values
    - **Answer: A**
49. **What does 'Early Stopping' help with in training models?**
    - A) Preventing overfitting by stopping training when performance on validation data starts to degrade
    - B) Increasing the learning rate
    - C) Reducing the number of features
    - D) Optimizing hyperparameters
    - **Answer: A**
50. **What does the term 'Epoch' refer to in training neural networks?**
    - A) One complete pass through the entire training dataset
    - B) The number of layers in a neural network
    - C) The size of a mini-batch
    - D) The learning rate of the model
    - **Answer: A**