NPTEL Data Analytics with Python - Top 30 MCQ Questions with Answers (2025)

1. Which of the following methods is used to handle missing values in a Pandas DataFrame?

A) df.na_values()

B) df.fillna()

C) df.replace_null()

D) df.handle_missing()

Answer: B) df.fillna()

2. In NumPy, which of the following operations demonstrates vectorization?

A) for i in range(len(arr)): arr[i] = arr[i] * 2

B) arr * 2

C) np.multiply_elements(arr, 2)

D) array.prod(arr, 2)

Answer: B) arr * 2

3. Which of the following is NOT a supervised learning algorithm?

A) Linear Regression

B) Random Forest

C) K-means Clustering

D) Support Vector Machines

Answer: C) K-means Clustering

4. Which Python library is primarily used for data manipulation and analysis?

A) NumPy

B) Pandas

C) Matplotlib

D) Scikit-learn

Answer: B) Pandas

5. The F1-score is the harmonic mean of:

A) Accuracy and Recall

B) Precision and Recall

C) Specificity and Sensitivity

D) Accuracy and Precision

Answer: B) Precision and Recall

6. In Python, which visualization library is built on top of Matplotlib and provides a high-level interface?

A) Plotly

B) Bokeh

C) Seaborn

D) ggplot

Answer: C) Seaborn

7. Which technique is commonly used to avoid overfitting in machine learning models?

A) Feature Engineering

B) Regularization

C) Normalization

D) Vectorization

Answer: B) Regularization

8. The process of converting categorical variables into numerical variables is known as:

A) Normalization

B) Standardization

C) Feature scaling

D) Encoding

Answer: D) Encoding

9. Which of the following metrics is NOT suitable for evaluating a regression model?

A) Mean Squared Error

B) R-squared

C) Mean Absolute Error

D) F1-score

Answer: D) F1-score

10. In K-means clustering, what criterion is often used to determine the optimal number of clusters?

```
A) Silhouette score
B) Confusion matrix
C) ROC curve
D) Precision-recall curve

**Answer: A) Silhouette score**
```

11. Which of the following is used for dimensionality reduction?

```
A) Random Forest
B) Principal Component Analysis
C) Logistic Regression
D) Gradient Boosting

**Answer: B) Principal Component Analysis**
```

12. What does the train_test_split function in scikit-learn do?

```
A) Splits data into training and testing sets
B) Splits algorithms into training and testing phases
C) Splits features and target variables
D) Splits data into clusters

**Answer: A) Splits data into training and testing sets**
```

13. Which of the following is NOT a type of join operation in Pandas?

```
A) Inner join
B) Outer join
C) Complex join
D) Left join

**Answer: C) Complex join**
```

14. The relationship between bias and variance in machine learning models is:

```
A) High bias often leads to overfitting
B) High variance often leads to underfitting
C) High bias often leads to underfitting
D) Bias and variance are unrelated concepts

**Answer: C) High bias often leads to underfitting**
```

15. Which of the following is NOT a method for handling imbalanced datasets?

```
A) Oversampling
B) Undersampling
C) SMOTE
D) Vectorization

**Answer: D) Vectorization**
```

16. In time series analysis, which technique is used to make a non-stationary series stationary?

```
A) Moving average
B) Differencing
C) Feature scaling
D) One-hot encoding

**Answer: B) Differencing**
```

17. Which ensemble learning method builds multiple decision trees and merges their predictions?

```
A) AdaBoost
B) Random Forest
C) Gradient Boosting
D) XGBoost

**Answer: B) Random Forest**
```

18. What does the term "bag of words" refer to in text analytics?

```
A) A collection of stop words
B) A representation where text is represented as word frequencies
C) A method for word embedding
D) A technique for handling missing words

**Answer: B) A representation where text is represented as word frequencies**
```

19. Which of these is NOT a hyperparameter of a Random Forest model?

```
A) Number of trees
B) Maximum depth
C) Feature coefficients
D) Minimum samples per leaf

**Answer: C) Feature coefficients**
```

20. In collaborative filtering for recommendation systems, what information is primarily used?

```
A) User demographics
B) Item features
C) User-item interaction patterns
D) Content descriptors

**Answer: C) User-item interaction patterns**
```

21. Which of the following is NOT a distance metric used in clustering algorithms?

```
A) Euclidean distance
B) Manhattan distance
C) Chebyshev distance
D) Gradient distance

**Answer: D) Gradient distance**
```

22. What does the value of k represent in k-fold cross-validation?

```
A) Number of features to select
B) Number of partitions the data is split into
C) Number of iterations for the algorithm
D) Number of clusters to form

**Answer: B) Number of partitions the data is split into**
```

23. Which of the following is used to visualize the distribution of a continuous variable?

```
A) Bar plot
B) Histogram
C) Scatter plot
D) Box plot

**Answer: B) Histogram**
```

24. The ROC curve plots:

```
A) Precision vs Recall
B) True Positive Rate vs False Positive Rate
C) Accuracy vs Threshold
D) Error vs Number of iterations

**Answer: B) True Positive Rate vs False Positive Rate**
```

25. Which of the following scaling methods ensures the resulting distribution has a mean of 0 and standard deviation of 1?

```
A) Min-Max scaling
B) Robust scaling
C) Standardization (Z-score normalization)
D) Log transformation

**Answer: C) Standardization (Z-score normalization)**
```

26. In Python, which function is used to find correlation between variables in a DataFrame?

```
A) df.correlation()
B) df.corr()
C) df.covariance()
D) df.relation()

**Answer: B) df.corr()**
```

27. Which of the following is NOT a common activation function in neural networks?

```
A) ReLU
B) Sigmoid
C) Tanh
D) Logarithmic

**Answer: D) Logarithmic**
```

28. What does LSTM stand for in the context of neural networks?

```
A) Long Short-Term Memory
B) Linear Standard Training Method
C) Logarithmic Statistical Temporal Model
D) Learning System Training Module

**Answer: A) Long Short-Term Memory**
```

29. Which of the following methods is used for feature selection in machine learning?

```
A) Forward Selection
B) Backward Elimination
C) Recursive Feature Elimination
D) All of the above

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

30. What problem does the "curse of dimensionality" refer to?

A) Difficulty in visualizing high-dimensional data

B) Issues with storing large datasets

C) The fact that data becomes sparse in high-dimensional spaces

D) The computational complexity of working with many features

Answer: C) The fact that data becomes sparse in high-dimensional spaces