Public Library · FBISE / NBF

Computer Science Grade 12 — Chapter 4

Data and Analysis

by Public Library

Independent practice MCQs from the supplied Class 12 Computer Science notes pack. Total: 50 questions.

A. Important / Main Topics

1. Which of the following is numerical data?

(A) Age
(B) Gender
(C) Color
(D) True/False

Correct Answer: A

Explanation: Age is numerical (quantitative) data because it represents a numerical value.

2. Which type of data has only two possible values?

(A) Numerical
(B) Boolean
(C) Categorical
(D) Text

Correct Answer: B

Explanation: Boolean — this correctly answers the question.

3. Which type of data is organized in rows and columns?

(A) Unstructured
(B) Semi-structured
(C) Structured
(D) Text

Correct Answer: C

Explanation: Structured data is organized in the form of rows and columns.

4. Which of the following is an example of unstructured data?

(A) Spreadsheet
(B) Database
(C) XML
(D) Video

Correct Answer: D

Explanation: Videos are unstructured data because they have no predefined structure.

5. Data visualization is best defined as:

(A) The graphical representation of information and data
(B) The process of collecting data
(C) The process of storing data
(D) The process of deleting data

Correct Answer: A

Explanation: Data visualization is the graphical representation of information and data.

6. Which Python libraries are used for data visualization?

(A) NumPy and Pandas
(B) Matplotlib and Seaborn
(C) TensorFlow and Keras
(D) Django and Flask

Correct Answer: B

Explanation: Matplotlib and Seaborn are Python libraries used for data visualization.

7. Data storytelling is best defined as:

(A) Writing a story about data
(B) Collecting data for a story
(C) Using visuals to communicate insights from data
(D) Deleting data

Correct Answer: C

Explanation: Data storytelling uses visuals to communicate insights, combining charts, graphs, and narratives.

8. Which of the following is a measure of central tendency?

(A) Range
(B) Standard deviation
(C) Variance
(D) Mean

Correct Answer: D

Explanation: Mean is a measure of central tendency, along with median and mode.

9. The mean of a dataset is:

(A) The average value of the data set
(B) The middle value when data is ordered
(C) The most frequent value
(D) The difference between highest and lowest values

Correct Answer: A

Explanation: The mean is the average value of the data set.

10. The median of a dataset is:

(A) The average value
(B) The middle value when data is ordered
(C) The most frequent value
(D) The difference between highest and lowest values

Correct Answer: B

Explanation: The median is the middle value when data is ordered.

11. The mode of a dataset is:

(A) The average value
(B) The middle value
(C) The most frequent value in the data
(D) The difference between highest and lowest values

Correct Answer: C

Explanation: The mode is the most frequent value in the data.

12. The range of a dataset is:

(A) The average value
(B) The middle value
(C) The most frequent value
(D) The difference between the highest and lowest values

Correct Answer: D

Explanation: The range is the difference between the highest and lowest values.

13. Which of the following is a key element of a data story?

(A) Relatable character
(B) Random numbers
(C) Complex equations
(D) Technical jargon

Correct Answer: A

Explanation: Relatable character — this correctly answers the question.

14. Hypothesis formulation is best defined as:

(A) Collecting data
(B) Making an educated guess about a relationship between variables
(C) Analyzing data
(D) Deleting data

Correct Answer: B

Explanation: Hypothesis formulation is the process of making an educated guess or assumption about a relationship between variables.

15. The null hypothesis (H□) states that:

(A) There is a significant effect or relationship
(B) The data is invalid
(C) There is no effect or relationship
(D) The hypothesis is proven

Correct Answer: C

Explanation: The null hypothesis states that there is no effect or relationship between variables.

16. How does continuous learning affect a machine learning model?

(A) It reduces the model's accuracy over time
(B) It makes the model slower to process data
(C) It simplifies the model's decision-making process
(D) It helps the model improve its performance by adapting to new data

Correct Answer: D

Explanation: It helps the model improve its performance by adapting to new data — this correctly answers the question.

17. What is feature engineering?

(A) Transforming and selecting relevant data features
(B) Collecting raw data
(C) Testing the model's accuracy
(D) Calculating the significance of the model

Correct Answer: A

Explanation: Transforming and selecting relevant data features — this correctly answers the question.

18. What is the main advantage of machine learning over rule-based algorithms?

(A) It requires fewer resources
(B) It can learn from data and adapt
(C) It is simpler to implement
(D) It does not require data

Correct Answer: B

Explanation: Machine learning can learn from data and adapt over time, unlike rule-based algorithms.

19. What is overfitting in machine learning?

(A) A model that is too simple
(B) A model that performs well on all data
(C) A model that learns the training data too well
(D) A model that does not learn at all

Correct Answer: C

Explanation: Overfitting happens when a model learns the training data too well, including noise, and performs poorly on new data.

20. What is underfitting in machine learning?

(A) A model that is too complex
(B) A model that learns the training data too well

(C) A model that performs well on all data
(D) A model that is too simple and fails to capture patterns

Correct Answer: D

Explanation: Underfitting occurs when a model is too simple and fails to capture underlying patterns, performing poorly on all data.

21. What is the test-train split used for?

(A) Dividing data into training and test portions
(B) Splitting the algorithm into two parts
(C) Dividing the model into two layers
(D) Splitting the output into two categories

Correct Answer: A

Explanation: A test-train split divides data into training (main) and test (smaller) portions for model evaluation.

22. What is supervised learning?

(A) Learning from unlabeled data
(B) Learning from labeled data
(C) Learning through rewards and penalties
(D) Learning without data

Correct Answer: B

Explanation: Supervised learning uses labeled data, where the correct output is provided for each input.

23. What is unsupervised learning?

(A) Learning from labeled data
(B) Learning through rewards and penalties
(C) Learning from unlabeled data
(D) Learning without data

Correct Answer: C

Explanation: Learning from unlabeled data — this correctly answers the question.

24. What is reinforcement learning?

(A) Learning from labeled data
(B) Learning from unlabeled data
(C) Learning without data
(D) Learning through rewards and penalties

Correct Answer: D

Explanation: Reinforcement learning involves a model that learns by interacting with an environment, receiving feedback in the form of rewards or penalties.

25. What is the formula for accuracy?

(A) Correct Predictions / Total Predictions
(B) True Positives / (True Positives + False Positives)
(C) Total Predictions / Correct Predictions
(D) False Positives / True Positives

Correct Answer: A

Explanation: Accuracy = Correct Predictions / Total Predictions.

B. Less Important / Secondary Topics

26. What is the formula for precision?

(A) True Positives / (True Positives + False Positives)
(B) Correct Predictions / Total Predictions
(C) False Positives / True Positives
(D) Total Predictions / Correct Predictions

Correct Answer: A

Explanation: True Positives / (True Positives + False Positives) — this correctly answers the question.

27. Which of the following is a measure of dispersion?

(A) Mean
(B) Standard deviation
(C) Median
(D) Mode

Correct Answer: B

Explanation: Standard deviation is a measure of dispersion, along with range and variance.

28. What does frequency distribution show?

(A) The average value
(B) The middle value
(C) How often each value occurs
(D) The difference between highest and lowest

Correct Answer: C

Explanation: Frequency distribution shows how often each value occurs in the data set.

29. Which measure is most affected by outliers?

(A) Median
(B) Mode
(C) Range
(D) Mean

Correct Answer: D

Explanation: The mean is most affected by outliers because it uses all values in its calculation.

30. Which measure is least affected by outliers?

(A) Median
(B) Mean
(C) Standard deviation
(D) Variance

Correct Answer: A

Explanation: The median is least affected by outliers because it only depends on the middle value.

31. What is data storytelling used for?

(A) To delete data
(B) To communicate insights from data
(C) To store data
(D) To encrypt data

Correct Answer: B

Explanation: Data storytelling uses visuals and narrative to communicate insights from data.

32. In a data story, what is the "call to action"?

(A) The data itself
(B) The character in the story
(C) A clear suggestion for what the audience should do next
(D) The chart used in the story

Correct Answer: C

Explanation: The call to action is a clear suggestion for what the audience should do next.

33. What is the purpose of formulating questions in data analysis?

(A) To make the analysis longer
(B) To eliminate the need for data
(C) To make the data more complex
(D) To guide data exploration and analysis

Correct Answer: D

Explanation: Formulating questions guides data exploration and analysis, ensuring the analysis is meaningful and directed.

34. What is the first step in the data analysis process?

(A) Formulating questions
(B) Identifying relevant data sets
(C) Evaluating data quality
(D) Refining questions

Correct Answer: A

Explanation: The first step is to formulate questions that guide the analysis.

35. Why is data quality important in data analysis?

(A) It makes the data more complex
(B) It ensures data accuracy and relevance
(C) It deletes data
(D) It stores data

Correct Answer: B

Explanation: Data quality ensures that the data is accurate and relevant, leading to reliable insights.

36. A model shows 98% accuracy, but precision and recall are both below 60%. What is the most likely issue?

(A) The model is underfitting
(B) The model is overfitting
(C) The data is imbalanced
(D) Labels are missing in training data

Correct Answer: C

Explanation: The data is imbalanced — this correctly answers the question.

37. A retail analyst wants to segment customers to discover hidden shopping patterns without any labeled categories. Which approach is most appropriate?

(A) Logistic regression
(B) Decision trees
(C) Linear regression
(D) Clustering

Correct Answer: D

Explanation: Clustering is an unsupervised learning technique that groups similar data points without labeled categories.

38. A machine learning model performs very well on training data but poorly on test data. What is the most likely cause?

(A) Overfitting
(B) Underfitting
(C) Data imbalance
(D) Missing labels

Correct Answer: A

Explanation: Overfitting occurs when a model learns the training data too well, including noise, and performs poorly on new, unseen data.

39. A company wants to predict house prices based on features like area, number of bedrooms, and location. Which type of learning is most appropriate?

(A) Unsupervised learning
(B) Supervised learning (regression)
(C) Reinforcement learning
(D) Clustering

Correct Answer: B

Explanation: Predicting continuous values like house prices is a regression problem, a type of supervised learning.

40. A researcher formulates a hypothesis that "increased study hours lead to better exam scores." What type of relationship does this hypothesis propose?

(A) Correlation
(B) Random
(C) Cause-and-effect
(D) No relationship

Correct Answer: C

Explanation: The hypothesis proposes a cause-and-effect relationship between study hours (cause) and exam scores (effect).

41. A study finds a p-value of 0.02 for the difference in plant height between fertilized and non-fertilized plants. What does this suggest?

(A) The difference is likely due to chance
(B) The data is invalid
(C) The fertilizer had no effect
(D) The difference is unlikely to be due to chance

Correct Answer: D

Explanation: A small p-value (0.02) suggests the difference is unlikely due to chance, providing evidence that the fertilizer had an effect.

42. A p-value of 0.5 is found in a study. What does this suggest?

(A) Weak evidence against the null hypothesis; the result could be due to chance
(B) Strong evidence against the null hypothesis
(C) The null hypothesis is proven true
(D) The alternative hypothesis is proven true

Correct Answer: A

Explanation: Weak evidence against the null hypothesis; the result could be due to chance — this correctly answers the question.

43. A company tests two versions of a website (A and B) to see which one leads to more sales. Which method is most appropriate?

(A) Clustering
(B) A/B testing
(C) Regression
(D) Qualitative testing

Correct Answer: B

Explanation: A/B testing compares two versions of a system to see which one performs better.

44. In the context of hypothesis testing, what is a Type I error?

(A) Failing to reject a false null hypothesis
(B) Rejecting a false null hypothesis
(C) Rejecting a true null hypothesis
(D) Failing to reject a true null hypothesis

Correct Answer: C

Explanation: A Type I error occurs when a true null hypothesis is incorrectly rejected (false positive).

45. In the context of hypothesis testing, what is a Type II error?

(A) Rejecting a true null hypothesis
(B) Rejecting a false null hypothesis
(C) Failing to reject a true null hypothesis
(D) Failing to reject a false null hypothesis

Correct Answer: D

Explanation: A Type II error happens when a false null hypothesis is not rejected (false negative).

46. What is power in hypothesis testing?

(A) The ability to detect an effect if one exists
(B) The probability of a Type I error
(C) The probability of a Type II error
(D) The average of the data

Correct Answer: A

Explanation: The ability to detect an effect if one exists — this correctly answers the question.

47. A larger sample size generally:

(A) Decreases power
(B) Increases power and reliability
(C) Has no effect on power
(D) Deletes data

Correct Answer: B

Explanation: A larger sample size generally Increases power and reliability.

48. A researcher wants to reduce the chance of a Type I error. What should they do?

(A) Increase the significance level (α)
(B) Increase the sample size
(C) Decrease the significance level (α)
(D) Decrease the sample size

Correct Answer: C

Explanation: Decrease the significance level (α) — this correctly answers the question.

49. A researcher wants to increase the power of a hypothesis test. Which of the following would help?

(A) Decrease the sample size
(B) Increase the significance level only
(C) Decrease the effect size
(D) Increase the sample size

Correct Answer: D

Explanation: Increase the sample size — this correctly answers the question.

50. In the coffee and productivity example, if the p-value is less than the significance level (α = 0.05), what is the correct decision?

(A) Reject H□ and accept H□
(B) Fail to reject H□
(C) Accept H□
(D) Ignore the results

Correct Answer: A

Explanation: If the p-value is less than the significance level, the null hypothesis is rejected, and the alternative hypothesis is accepted.