(7 pages)

Reg. No. : .....

Code No. : 7900 **Sub. Code : WCSM 33**

M.Sc. (CBCS) DEGREE EXAMINATION, NOVEMBER 2024.

Third Semester

Computer Science — Core

DATA SCIENCE AND ANALYTICS

(For those who joined in July 2023 onwards)

Time : Three hours Maximum : 75 marks

PART A — (15 × 1 = 15 marks)

Answer ALL questions.

Choose the correct answer :

1. ———— is a multidisciplinary which involves extraction of knowledge from large volumes of data that are structured or unstructured.
   (a) Data analysis
   (b) Data science
   (c) Descriptive analysis
   (d) Data mining

2. Which of the following is characteristic of processed data?
   (a) Data is not ready for analysis
   (b) All steps should be noted
   (c) Hard to use for data analysis
   (d) Soft to use for data analysis

3. Data that summarize all observations in a category are called ———— data.
   (a) frequency (b) summarized
   (c) raw (d) meta data

4. Which of the following approach should be used to ask data analysis question?
   (a) Find only one solution for particular problem
   (b) Find out the question which is to be answered
   (c) Find out answer from dataset without asking question
   (d) Find out answer from dataset without non-question

5. Which of the following is another name for raw data?
   (a) destination data (b) eggy data
   (c) secondary (d) machine learning

6. Which of the following language is used in data science?

(a) R (b) C

(c) C++ (d) Ruby

7. Which of the following function gives information about top level data?

(a) head (b) tail

(c) summary (d) layer

8. What is 'Big Data'?

(a) Large volumes of complex data

(b) The study of algorithms

(c) Data that is easy to process

(d) Small and simple datasets

9. What does 'Clustering' mean in the context of machine learning?

(a) Dividing the dataset into sets

(b) Grouping similar items together

(c) Predicting the outcome for new data

(d) Reducing the dimensionality of data

10. What is 'Principal component analysis' used for?

(a) To increase the size of data

(b) To decrease the computational complexity of data

(c) For data augmentation

(d) Dimensionality reduction

11. Which of the following is statistical boosting based on additive logistic regression?

(a) gamBoost (b) gbm

(c) ada (d) mboost

12. Which of the following focuses on the discovery of (previously) unknown properties on the data?

(a) Data mining (b) Big data

(c) Data wrangling (d) Machine learning

13. Which of the following technique comes under practical machine learning?

(a) Decision tree (b) Forecasting

(c) Data visualisation (d) Clustering

14. Which of the following step is performed by data scientist after acquiring the data?

(a) Data replication (b) Data integration

(c) Data forecasting (d) Data cleaning

15. Which of the following data is put into a formula to produce commonly accepted results?

(a) Raw (b) Processed

(c) Synchronized (d) Asynchronies

PART B — (5 × 4 = 20 marks)

Answer ALL questions, choosing either (a) or (b).
Each answer should not exceed 250 words.

16. (a) Distinguish between the structured data and unstructured data.

Or

(b) Describe the exploratory data analysis in data science.

17. (a) Elaborate the data preparation in data analytics lifecycle.

Or

(b) Summarize the advantages of NoSQL.

18. (a) Decide the steps to data import and export using R.

Or

(b) Highlight the visualizing a single variable with example.

19. (a) Explain the need of decision trees in R.

Or

(b) Discuss the basic concept of Bayes theorem.

20. (a) Identify the use of deep learning in data science.

Or

(b) Recall the process of logistic regression.

PART C — (5 × 8 = 40 marks)

Answer ALL questions, choosing either (a) or (b)
Each answer should not exceed 600 words.

21. (a) Examine the big data ecosystem and data science.

Or

(b) Determine the different types of machine learning.

22. (a) Appraise the statistical methods for evaluation in data analytics.

Or

(b) Illustrate the implementation of Hadoop ecosystem.

23. (a) Compare and construct the data exploration and presentation.

Or

(b) Elaborate the dirty data in exploratory data analysis.

24. (a) Discuss the steps to evaluating a decision tress.

Or

(b) Assume the implementation of Nai've Bayes classifier.

25. (a) Demonstrate the use of clustering in artificial intelligence.

Or

(b) Conclude the process of linear regression with example.
