

CLASS: XII
ARTIFICIAL INTELLIGENCE (SUBJECT CODE - 843)
UNIT 2: Data Science Methodology

◆ **Section 1: Basics of Data Science Methodology**

1. Data Science Methodology provides a framework for:

- A) Random experimentation
- B) Systematic problem solving
- C) Entertainment
- D) Data storage only

Answer: B

2. How many steps are there in the Data Science Methodology proposed by John Rollins?

- A) 5
- B) 7
- C) 10
- D) 12

Answer: C

3. The first stage in Data Science Methodology is:

- A) Data Collection
- B) Business Understanding
- C) Modelling
- D) Evaluation

Answer: B

4. Which tool is used to understand the problem deeply in Business Understanding stage?

- A) Confusion Matrix
- B) 5W1H Problem Canvas
- C) Decision Tree
- D) K-Means

Answer: B

◆ **Section 2: Analytic Approach**

5. Which approach answers “How much or how many?”

- A) Classification
- B) Regression
- C) Clustering
- D) Recommendation

Answer: B

6. Identifying categories of data belongs to:

- A) Regression
- B) Classification
- C) Clustering
- D) Anomaly Detection

Answer: B

7. Detecting unusual patterns in data is called:

- A) Classification
- B) Regression
- C) Clustering
- D) Anomaly Detection

Answer: D

◆ Section 3: Types of Analytics

8. Which analytics explains “what happened”?

- A) Diagnostic
- B) Predictive
- C) Descriptive
- D) Prescriptive

Answer: C

9. Which analytics explains “why it happened”?

- A) Descriptive
- B) Diagnostic
- C) Predictive
- D) Prescriptive

Answer: B

10. Which analytics predicts future outcomes?

- A) Descriptive
- B) Diagnostic
- C) Predictive
- D) Prescriptive

Answer: C

11. Which analytics suggests actions to achieve goals?

- A) Predictive
- B) Diagnostic
- C) Prescriptive
- D) Descriptive

Answer: C

◆ Section 4: Data Requirements & Collection

12. Data organized in tables is called:

- A) Unstructured data
- B) Semi-structured data
- C) Structured data
- D) Raw data

Answer: C

13. Emails are an example of:

- A) Structured data
- B) Semi-structured data
- C) Unstructured data
- D) Numeric data

Answer: B

14. Data collected firsthand is called:

- A) Secondary data
- B) Primary data
- C) Raw data
- D) Processed data

Answer: B

15. Which of the following is a secondary data source?

- A) Surveys
- B) Interviews
- C) Books and websites
- D) Experiments

Answer: C

◆ Section 5: Data Understanding & Preparation

16. Checking whether data represents the problem is done in:

- A) Data Collection
- B) Data Understanding
- C) Data Preparation
- D) Modelling

Answer: B

17. Removing duplicate values is part of:

- A) Data Collection
- B) Data Preparation

- C) Modelling
- D) Deployment

Answer: B

18. Feature Engineering refers to:

- A) Collecting data
- B) Cleaning data
- C) Creating new variables
- D) Deleting data

Answer: C

◆ Section 6: Modelling & Evaluation

19. A dataset used to train a model is called:

- A) Test set
- B) Validation set
- C) Training set
- D) Sample set

Answer: C

20. Which model describes data without making predictions?

- A) Predictive model
- B) Descriptive model
- C) Classification model
- D) Regression model

Answer: B

21. Which metric measures correct predictions over total predictions?

- A) Precision
- B) Recall
- C) Accuracy
- D) F1 Score

Answer: C

22. Precision is defined as:

- A) $TP / (TP + FN)$
- B) $TP / (TP + FP)$
- C) $TN / Total$
- D) $FP / Total$

Answer: B

23. Recall is defined as:

- A) $TP / (TP + FN)$
- B) $TP / (TP + FP)$
- C) TN / Total
- D) FP / Total

Answer: A

24. F1 Score is:

- A) Average of accuracy and precision
- B) Harmonic mean of precision and recall
- C) Sum of precision and recall
- D) Difference of precision and recall

Answer: B

◆ Section 7: Model Validation

25. Train-test split divides data into:

- A) 3 parts
- B) 4 parts
- C) 2 parts
- D) 5 parts

Answer: C

26. Common train-test split ratio is:

- A) 90:10
- B) 80:20
- C) 50:50
- D) 60:40

Answer: B

27. K-Fold Cross Validation improves:

- A) Speed
- B) Data size
- C) Accuracy estimation
- D) Storage

Answer: C

◆ Section 8: Evaluation Metrics (Regression)

28. MAE stands for:

- A) Mean Average Error
- B) Mean Absolute Error
- C) Maximum Absolute Error

D) Mean Actual Error

Answer: B

29. RMSE is preferred because:

- A) It is smaller
- B) It uses less data
- C) It is in same unit as target variable
- D) It ignores errors

Answer: C

30. MSE is:

- A) Average of absolute errors
- B) Average of squared errors
- C) Sum of errors
- D) Difference of errors

Answer: B

31. Which is the hardest stage in the foundational methodology of Data Science?

- a. **Business Understanding** b. Data collection
- c. Modelling d. Evaluation

32. Business Sponsors defines the problem and project objectives from a _____ perspective.

- a. Economic b. Feedback c. **Business** d. Data Collection

33. Match the following and choose the correct options:

- i. Descriptive approach A. Statistical Analysis
 - ii. Diagnostic approach B. Current Status
 - iii. Predictive approach C. How to solve it?
 - iv. Prescriptive approach D. Probabilities of action
- a. (i)—A , (ii)—B, (iii) – C , (iv)—D b. **(i)—B , (ii)—A, (iii) – D , (iv)—C**
- c. (i)—D , (ii)—B, (iii) – A , (iv)—C d. (i)—A , (ii)—C, (iii) – B , (iv)—D

34. Arrange the following statements in order

- i: Gaps in data will be identified and plans to fill/make substitutions will have to be made
 - ii: Decisions are made whether the collection requires more data or not
 - iii: Descriptive statistics and visualization is applied to dataset
 - iv: Identify the necessary data content, formats and sources
- a. i,ii,iii,iv b. **iv,ii,iii,i** c. i,iii,ii,iv d. ii,i,iii,iv

35. Data Modelling focuses on developing models that are either _____ or _____

- a. Supervised, Unsupervised b. **Predictive, Descriptive**
- c. Classification, Regression d. Train-test split, Cross Validation

36. Statement 1- There is no optimal split percentage

Statement 2- The most common split percentage between training and testing data is 20%-80%

- a. Statement 1 is true Statement 2 is false
- b. Statement 2 is true Statement 1 is false
- c. Both Statement 1 and 2 are true
- d. Both Statement 1 and 2 are false

37. Train-test split function is imported from which Python module?

- a. [sklearn.model_selection](#) b. sklearn.ensemble
- c. sklearn.metrics d. sklearn.preprocessing

38. Identify the incorrect statement:

- i. cross-validation gives a more reliable measure of your model's quality
 - ii. cross-validation takes short time to run
 - iii. cross-validation gets multiple measures of model's quality
 - iv. cross-validation is preferred with small data
- a. ii and iii b. iii only c. [ii only](#) d. ii, iii and iv

39. Identifying the necessary data content, formats and sources for initial data collection is done in which step of Data Science methodology? a. [Data requirements](#) b. Data Collection c. Data Understanding d. Data Preparation

40. Data sets are available online. From the given options, which one does not provide online data?

- a. UNICEF b. WHO c. Google d. [Edge](#)

41. A _____ set is a set of historical data in which outcomes are already known.

- a. [Training set](#) b. Test set c. Validation set d. Evaluation set

42. _____ data set is used to evaluate the fit machine learning model.

- a. Training set b. [Test set](#) c. Validation set d. Evaluation set

43. `x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.2)`

From the above line of code, identify the training data set size

- a. 0.2 b. [0.8](#) c. 20 d. 80

44. In k-fold cross validation, what does k represent?

- a. number of subsets b. number of experiments c. number of folds d. [all of the above](#)

45. Identify the correct points regarding MSE given below:

- i. MSE is expanded as Median Squared Error
 - ii. MSE is standard deviation of the residuals
 - iii. MSE is preferred with regression
 - iv. MSE penalize large errors more than small errors
- a. i and ii b. ii and iii c. [iii and iv](#) d. ii, iii and iv

B. Short Answer Questions

1. How many steps are there in Data Science Methodology? Name them in order.
2. What do you mean by Feature Engineering?
3. Data is collected from different sources. Explain the different types of sources with example.
4. Which step of Data Science Methodology is related to constructing the data set? Explain.
5. Write a short note on the steps done during Data Preparation.
6. Differentiate between descriptive modelling and predictive modelling.
7. Explain the different metrics used for evaluating Classification models.
8. Is Feedback a necessary step in Data Science Methodology? Justify your answer.
9. Write a comparative study on train-test split and cross validation.
10. Why is model validation important?

C. Long Answer Questions

1. Explain the procedure of k-fold cross validation with suitable diagram.

2. Data is the main part of any project. How will you find the requirements of data, collect it, understand the data and prepare it for modelling?

3. Given a confusion matrix, calculate Precision, Recall, F1 score and Accuracy.

$$\begin{bmatrix} \text{TN} = 90 & \text{FP} = 20 \\ \text{FN} = 15 & \text{TP} = 75 \end{bmatrix}$$

D. Competency Based Questions

1. A leading investment firm aims to improve their client portfolio management system. They

want to know whether Artificial Intelligence could be used to better understand clients' investment preferences and risk tolerance levels. Which stage of Data Science methodology can you relate this to?

2. An Online Learning Platform has implemented a recommendation system to suggest personalized courses to users. They need to assess the effectiveness and accuracy of this system. Which stage of Data Science methodology can you relate this to?