

Total No. of Printed Pages—8

SEC45CSC102(T)25

2nd Semester Exam., 2026 (May) (FYUP)

SEC-2 (COMPUTER SCIENCE)

[Foundation of Data Science]

Full Marks : 70

Pass Marks : 28

Time : 3 hours

General Instructions :

1. All questions are compulsory.
2. Marks for each question are indicated in the figure.

A. Choose the correct answer and rewrite them :

1×10=10

1. Which of the following is used to summarize and describe the main features of a dataset?

- (A) Inferential statistics
- ~~(B)~~ Descriptive statistics
- (C) Predictive analytics
- (D) Prescriptive analytics

26Q/1364

(Turn Over)

(2)

2. The main objective of ordinary least squares method is to
- (A) maximize correlation
 - ☒ (B) minimize sum of squared residuals
 - (C) maximize accuracy
 - (D) minimize variance
3. Logistic regression is primarily used for
- (A) predicting continuous values
 - ☒ (B) binary/multi-class classification
 - (C) clustering data points
 - (D) finding association rules
4. Which is **not** a challenge for Big Data Analytics as per syllabus?
- (A) Volume
 - (B) Velocity
 - ☒ (C) Simplicity
 - (D) Variety
5. In active learning, the algorithm selects data points for labelling that
- ☒ (A) model is most confident about
 - ☒ (B) model is most uncertain/confused about
 - (C) are farthest from centre
 - (D) are selected randomly from dataset

(3)

6. Choose the correct combination from the following matrix :

Column—I (Technique)	Column—II (Learning Type)	Column—III (Output)
a. K-Means	1. Unsupervised	(i) Support, Confidence, Lift
b. Apriori	2. Supervised	(ii) Cluster centers + labels
c. SVM	3. Unsupervised	(iii) Dendrogram tree
d. Hierarchical Clustering	4. Supervised	(iv) Maximum margin hyperplane

☒ (A) a b c d
 1 1 4 3
 (ii) (i) (iv) (iii)

(B) a b c d
 2 3 1 4
 (iv) (iii) (i) (ii)

(C) a b c d
 3 4 2 1
 (iii) (ii) (i) (iv)

(D) a b c d
 4 2 3 1
 (i) (iv) (ii) (iii)

7. Choose the correct statements about descriptive statistics and hypothesis testing :
1. Mean is affected by outliers but median is not.
 2. In normal distribution, mean = median = mode.
 3. Type-II Error = Rejecting H_0 when H_0 is true.
 4. P-value < 0.05 means result is statistically significant.
 5. Variance can never be negative.
- (A) 1 and 2
(B) 2 and 3
☒ (C) 1, 2, 4 and 5
~~(D) 3 and 4~~
8. What is the correct sequence to build a decision tree?
- a. Calculate information gain/gini for each feature
 - b. Split dataset on feature with max gain
 - c. Repeat for child nodes until stopping condition
 - d. Select root node as the best feature
- ~~(A) a→d→b→c~~
(B) d→a→b→c
(C) b→a→d→c
(D) c→b→a→d

9. Statement I :
Hierarchical clustering does not require pre-specifying number of clusters.
- Statement II :
Elbow method plots inertia vs K to find optimal K in K-means.
- Choose the correct option.
- ☒ (A) Both Statement I and Statement II are correct.
(B) Statement I is correct, but Statement II is incorrect.
(C) Statement I is incorrect, but Statement II is correct.
(D) Both Statement I and Statement II are incorrect
10. Assertion (A) :
Standard deviation is a better measure of dispersion than range.
- Reason (R) :
Range is affected only by extreme values, while SD considers all data points.
- Choose the correct option.
- ☒ (A) Both (A) and (R) are true, and (R) explains (A).
(B) Both (A) and (R) are true, but (R) does not explain (A).
(C) (A) is true, but (R) is false
(D) (A) is false, but (R) is true.

(6)

B. Answer all the questions in approximately 30-40 words each : $2 \times 5 = 10$

11. What is hypothesis testing?
12. Differentiate between regression and classification with example of each.
13. A logistic model gives probability = 0.45 for "Churn = Yes". Business threshold = 0.5. What action company should take?
14. Define support and confidence in association rule mining.

Or

State two differences between K-means and hierarchical clustering.

15. What is reinforcement learning? Name two components of reinforcement learning.

Or

Differentiate among descriptive, predictive and prescriptive analytics.

C. Answer all the questions in around 80-100 words each : $5 \times 6 = 30$

16. Explain type 1 error and type 2 error with example.
17. What is K-nearest neighbours algorithm? Explain why it is called lazy learner.

(7)

18. Support vector machine gives value w into X plus b equals minus 1.5 for customer data. If value greater than 0 means class plus 1 and less than 0 means class minus 1, which class will customer get?

19. Explain support vector machine concept of margin and support vectors.

Or

Explain logistic regression and sigmoid function role in it.

20. Explain active learning and why it reduces labeling cost.

Or

What is feedback loop in reinforcement learning?

21. Read the passage carefully and write the answer to the questions from the passage :

"A bank built loan approval model. On training data model accuracy was 99 percent. On new customer data accuracy dropped to 55 percent. Data scientists checked and found the model was too complex with 50 features. It memorized all training examples including noise. This is high variance and overfitting problem. Bias of model was very low, 0.5. To fix it they used regularization technique which penalizes large feature values. After regularization training accuracy became 85 percent and

test accuracy became 80 percent. Gap reduced and model became stable."

Based on the above case, answer the following questions :

- (a) What problem occurred in bank model?
Give two evidences. 2
- (b) What solution was applied and what was the result after solution? 2
- (c) What was the bias value of original model? 1

D. Answer all questions within 180-200 words each : $10 \times 2 = 20$

22. Explain market basket analysis and concept of frequent itemset.

Or

Explain density based clustering concept and advantage over K-means.

23. A company has two machines. Machine A produces packets with average 500 gram and standard deviation 2 gram. Machine B produces packets with average 505 gram and standard deviation 15 gram. Customer wants packets close to 500 gram and consistent weight. Which machine is better? Justify using statistical concepts.

Or

Two hospitals test same medicine. Hospital X tests on 50 patients, Hospital Y on 5000 patients. Both get 60% success rate. Which result is more reliable? Justify using sample size concept.