

**SET-4****Series &RQPS/S****Q.P. Code 368/S**

Roll No.

--	--	--	--	--	--	--	--

Candidates must write the Q.P. Code on the title page of the answer-book.

- Please check that this question paper contains **11** printed pages.
- Please check that this question paper contains **21** questions.
- Q.P. Code given on the right hand side of the question paper should be written on the title page of the answer-book by the candidate.
- **Please write down the serial number of the question in the answer-book before attempting it.**
- 15 minute time has been allotted to read this question paper. The question paper will be distributed at 10.15 a.m. From 10.15 a.m. to 10.30 a.m., the students will read the question paper only and will not write any answer on the answer-book during this period.



DATA SCIENCE

*Time allowed : 2 hours**Maximum Marks : 50*

General Instructions :

- Please read the instructions carefully.*
- This question paper consists of **21** questions in **two** sections : **Section A** and **Section B**.*
- Section A has Objective Type Questions, whereas Section B contains Subjective Type Questions.*
- Out of the given (5 + 16 =) 21 questions, the candidate has to answer (5 + 10 =) 15 questions in the allotted (maximum) time of 2 hours.***
- All questions of a particular section must be attempted in the correct order.*



(vi) **Section A : Objective Type Questions (24 marks) :**

- (a) This section has **5** questions.
- (b) There is no negative marking.
- (c) Do as per the instructions given.
- (d) Marks allotted are mentioned against each question/part.

(vii) **Section B : Subjective Type Questions (26 marks) :**

- (a) This section has **16** questions.
- (b) A candidate has to do **10** questions.
- (c) Do as per the instructions given.
- (d) Marks allotted are mentioned against each question/part.

SECTION A

(Objective Type Questions)

(24 marks)

1. Answer any **4** out of the given **6** questions on Employability Skills. $4 \times 1 = 4$

- (i) Match the following personalities with their parameters, based on Five-Factor model.

(a) Openness	(i) Confident and easily make friends
(b) Agreeableness	(ii) Creative, curious and adventurous
(c) Extraversion	(iii) Accommodate themselves in any situation, cooperative and considerate

- (A) (a) - (i), (b) - (iii), (c) - (ii)
- (B) (a) - (ii), (b) - (i), (c) - (iii)
- (C) (a) - (ii), (b) - (iii), (c) - (i)
- (D) (a) - (iii), (b) - (i), (c) - (ii)



- (ii) _____ is the ability of an entrepreneur to do something, even when it is difficult.
- (A) Perseverance (B) Decisiveness
(C) Interpersonal skills (D) Organizational skills
- (iii) How many basic punctuation marks or signs are used in English language ?
- (A) 10 (B) 12
(C) 15 (D) 18
- (iv) In a spreadsheet, what does the equal sign (=) signify ?
- (A) It denotes a subtraction operation.
(B) It signifies the beginning of a comment.
(C) It represents a cell reference.
(D) It indicates the start of a formula or calculation.
- (v) *Assertion (A)* : Motivation that arises because of incentives or external rewards is known as extrinsic motivation.
Reason (R) : Extrinsic motivation is characterised by the influence of external factors such as rewards, recognition or punishments, which drive individuals to engage in specific behaviours or tasks.
- (A) Both Assertion (A) and Reason (R) are true and Reason (R) is the correct explanation for Assertion (A).
(B) Both Assertion (A) and Reason (R) are true, but Reason (R) is **not** the correct explanation for Assertion (A).
(C) Assertion (A) is true, but Reason (R) is false.
(D) Assertion (A) is false, but Reason (R) is true.
- (vi) What are green jobs primarily focused on ?
- (A) Maximising profits
(B) Environmental conservation and sustainability
(C) Industrial automation
(D) Manufacturing industry



2. Answer any 5 out of the given 6 questions.

5×1=5

(i) What is the primary goal of Exploratory Data Analysis (EDA) ?

- (A) To build predictive models
- (B) To summarise data for reporting
- (C) To perform hypothesis testing
- (D) To explore and understand data patterns

(ii) Match the following with respect to decision trees.

(a) Leaf node	(i) Outcome of test
(b) Internal node	(ii) Class label
(c) Branch	(iii) Question on choosing a particular class

- (A) (a) - (ii), (b) - (i), (c) - (iii)
- (B) (a) - (ii), (b) - (iii), (c) - (i)
- (C) (a) - (iii), (b) - (i), (c) - (ii)
- (D) (a) - (iii), (b) - (ii), (c) - (i)

(iii) State **True** or **False**.

Data usability is not a focus area of data governance.

(iv) KNN stands for _____.

- (A) Knowledgeable Neural Network
- (B) K-Means Nearest Neighbors
- (C) K-Nearest Neighbors
- (D) Kernelised Neural Network

(v) What term is used to describe regression methods that have a linear function relationship between dependent and independent variables ?

- (A) Linear regression models
- (B) Non-linear regression models
- (C) Exponential regression models
- (D) Polynomial regression models



(vi) State **True** or **False**.

Linear regression is more flexible than non-linear regression.

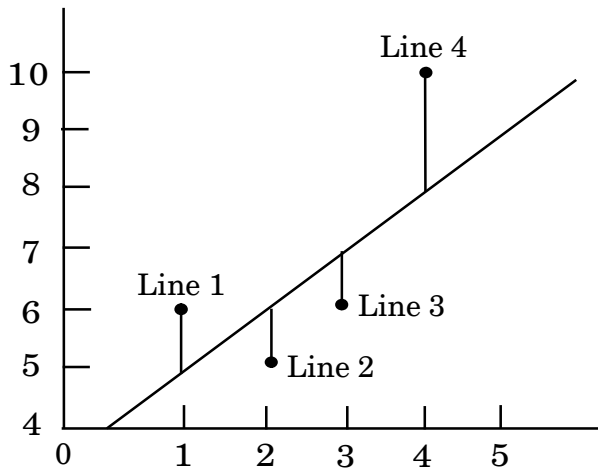
3. Answer any **5** out of the given **6** questions.

5×1=5

- (i) ABC Corporation is storing data from multiple sources for analysis and improving customer experiences. What should be their primary consideration regarding data storage ?
- (A) Storing all data indefinitely for future use.
 - (B) Implementing strict data storage rules to comply with regulations.
 - (C) Sharing the data openly with third-party companies.
 - (D) Storing data without any regard for privacy concerns.
- (ii) What are common graphical methods used for conducting bivariate analysis ?
- (A) Scatter plots and Counter plots
 - (B) Cluster analysis and Pair plots
 - (C) Line charts and Counter plots
 - (D) Scatter plots and Pair plots
- (iii) Decisions trees for regression are used when the target variable is :
- (A) Categorical
 - (B) Continuous
 - (C) Binary
 - (D) Textual
- (iv) Which algorithm is very sensitive to outliers in the dataset ?
- (A) Decision Trees
 - (B) Logistic Regression
 - (C) K-NN
 - (D) Random Forest



- (v) Consider the graph given below :



In the above graph, what do vertical lines (Line 1, Line 2, Line 3, and Line 4) represent ?

- (A) Lines of best fit
 - (B) Observed values
 - (C) Error or Residuals
 - (D) RSME values
- (vi) *Assertion (A)* : K-means clustering is an unsupervised machine learning technique.

Reason (R) : For K-means clustering, we need to have some data tagged with correct labels that can be used for training.

- (A) Both Assertion (A) and Reason (R) are correct and Reason (R) is the correct explanation of Assertion (A).
- (B) Both Assertion (A) and Reason (R) are correct, but Reason (R) is **not** the correct explanation for Assertion (A).
- (C) Assertion (A) is true, but Reason (R) is false.
- (D) Reason (R) is true, but Assertion (A) is false.



4. Answer any 5 out of the given 6 questions.

5×1=5

- (i) _____ was passed in the United States to protect healthcare information from fraud and theft.
- (A) General Data Protection Regulation
(B) Health Insurance Portability and Accountability Act
(C) Personal Data Protection Bill
(D) California Consumer Privacy Act
- (ii) A researcher is conducting a study to understand the joint influence of temperature, humidity and wind speed on energy consumption in a building. What type of analysis is most appropriate for this study ?
- (A) Inferential Analysis
(B) Multivariate Analysis
(C) Univariate Analysis
(D) Bivariate Analysis
- (iii) In case of a classification tree, the value or class of the terminal nodes after training is the _____ of the observation.
- (A) mean (B) mode
(C) median (D) sum
- (iv) *Statement I* : K-NN algorithm is sensitive to outliers.
Statement II : If there are outliers in the data, accuracy of K-NN algorithm increases.
- (A) Both statement I and statement II are correct
(B) Both statement I and statement II are incorrect
(C) Statement I is correct, but statement II is incorrect
(D) Statement II is correct, but statement I is incorrect



(v) Match the following with respect to Decision Trees.

(a) RMSE	(i) Predicting the value of dependent variable on the basis of independent variable
(b) MAE	(ii) Average magnitude of errors
(c) Linear Regression	(iii) Square root of variance of residuals

(A) (a) - (ii), (b) - (i), (c) - (iii)

(B) (a) - (iii), (b) - (ii), (c) - (i)

(C) (a) - (i), (b) - (ii), (c) - (iii)

(D) (a) - (i), (b) - (iii), (c) - (ii)

(vi) _____ learning helps to build better buyer persona profiles.

5. Answer any 5 out of the given 6 questions.

5×1=5

(i) State **True** or **False**.

Software products and data are always used for purposes that are good for society.

(ii) *Statement I* : K-NN uses all the training data while performing a classification operation.

Statement II : K-NN does not assume anything about distribution of data.

(A) Both statement I and statement II are correct

(B) Both statement I and statement II are incorrect

(C) Statement I is correct, but statement II is incorrect

(D) Statement II is correct, but statement I is incorrect

(iii) _____ techniques apply where no class of data is to be predicted.



- (iv) In linear regression, the graph of the relationship between the dependent and independent variables follows an equation that represents a _____.
- (A) straight line
 - (B) parabola
 - (C) linear equation
 - (D) curve
- (v) Which of the following is a real-world application where unsupervised learning techniques are commonly used ?
- (A) Rainfall prediction
 - (B) House price prediction
 - (C) Medical imaging
 - (D) Spam email detection
- (vi) _____ can be used to find the effect of age, weight and height on cholesterol levels in your body.
- (A) Correlation
 - (B) Linear regression
 - (C) Non-linear regression
 - (D) Multiple linear regression

SECTION B

(Subjective Type Questions)

(26 marks)

Answer any 3 out of the given 5 questions on Employability Skills. Answer each question in 20 – 30 words.

3×2=6

6. Define the following terms :

- (a) First generation entrepreneurs
- (b) Women entrepreneurs



7. Identify the type of sentence used in the statement as declarative, interrogative, exclamatory or imperative :
- (a) How are you ?
 - (b) It's very cold !
 - (c) Eat your food.
 - (d) I received a gift voucher.
8. Define stress. Write any two ways to manage stress.
9. What is the difference between a Formula Bar and Name Box in a spreadsheet ?
10. Explain the concept of global warming and its relationship with greenhouse gas emissions. Additionally, mention any one specific measure that is taken to reduce greenhouse gas emissions in the context of energy sources.

Answer any 4 out of the given 6 questions in 20 – 30 words each.

4×2=8

11. One major aspect of data privacy is that the individual is considered to be the sole owner of data.
Is the given statement correct with respect to data privacy ? Why/Why not ?
12. Define Exploratory Data Analysis (EDA). Give names of any two tools/methods that can be used to perform EDA.
13. List any two features of a Decision Tree.
14. Define Cross Validation. What is its use in data science ?
15. Explain the signification of the Root Mean Squared Error (RMSE) in linear regression. How does a smaller RMSE value reflect the quality of a regression model ?
16. Why do websites use recommendation engines along with unsupervised learning techniques ?



Answer any **3** out of the given **5** questions in 50 – 80 words each.

$3 \times 4 = 12$

- 17.** (a) Write a short note on PDP – Personal Data Protection Bill.
(b) Mention any two ethical guidelines that one must adhere to while dealing with data.
- 18.** Imagine you are tasked with conducting a research project involving a dataset that records the salary of employees in ABC institution. However, this dataset has not yet undergone the data cleaning process. Explain any four specific data cleaning steps you would implement to prepare the dataset for analysis.
- 19.** Differentiate between Regression tree and Classification tree. Also provide a labelled diagram of a simple decision tree.
- 20.** Define Multiple Linear Regression. Also, write and explain the formula of multiple linear regression.
- 21.** Explain any four real world applications of Unsupervised Learning.