



Series &RQPS

SET-4

Q.P. Code 368

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 candidates 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 :

- (i) *Please read the instructions carefully.*
- (ii) *This question paper consists of **21** questions in **two** Sections : Section – **A** & Section – **B**.*
- (iii) *Section – **A** has Objective Type Questions whereas Section – **B** contains Subjective Type Questions.*
- (iv) *Out of the given $(5 + 16 =)$ **21** questions, a candidate has to answer $(5 + 10 =)$ **15** questions in the allotted (maximum) time of **2** hours.*
- (v) *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 05 questions.*
 - (b) *Marks allotted are mentioned against each question/part.*
 - (c) *There is no **negative** marking.*
 - (d) *Do as per the instructions given.*
- (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)

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. 1
- | | | |
|------------------|-------|--|
| a. Extraversion | (i) | Self disciplined and take care of others before themselves |
| b. Neuroticism | (ii) | Talkative, confident and make any gathering lively |
| c. Consciousness | (iii) | Self doubt, depression and shyness. |
- (A) a – (ii), b – (iii), c – (i)
(B) a – (i), b – (ii), c – (iii)
(C) a – (iii), b – (ii), c – (i)
(D) a – (ii), b – (i), c – (iii)
- (ii) A _____ is a company that is in the first stage of its operations. 1
- (A) Corporation (B) Startup
(C) Multinational (D) Conglomerate
- (iii) Which non-verbal cue is not associated with active listening ? 1
- (A) Nodding in agreement
(B) Maintaining an open and welcoming body posture
(C) Checking your phone when the other person is speaking
(D) Making encouraging facial expressions
- (iv) **Statement (A)** : To sum the values in cells A1 through A10 in a spreadsheet, the formula used is "=SUM(A1:A10)". 1
- Statement (B)** : The SUM function in spreadsheet software is specifically designed to add up a range of cells.
- (A) Both (A) and (B) are true.
(B) Both (A) and (B) are false.
(C) (A) is true, but (B) is false.
(D) (A) is false, but (B) is true.



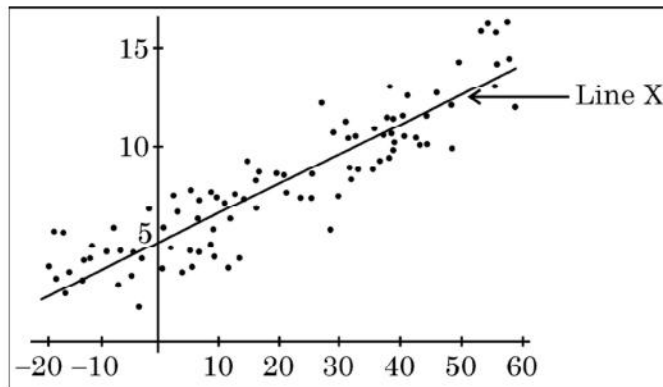
- (v) Why is setting SMART goals important in self-management ? 1
- (A) It makes goals too rigid.
- (B) It allows for vague and undefined goals.
- (C) It defines action based steps that one will follow to achieve the goal.
- (D) It discourages goal setting altogether.
- (vi) Which of the following is a benefit of green jobs ? 1
- (A) They do not support adaptation to climate change.
- (B) They control waste and pollution.
- (C) They increase greenhouse gas emissions.
- (D) They decrease efficiency of raw materials.
2. Answer any 5 out of the given 6 questions. 5 × 1 = 5
- (i) EDA stands for : 1
- (A) Exploratory Data Analysis
- (B) Efficient Data Accumulation
- (C) Extended Data Assessment
- (D) Essential Data Application
- (ii) **Assertion (A) :** Decision trees are versatile algorithms and are easy to visualize. 1
- Reasoning (R) :** Decision trees can only be used to map linear relationships.
- (A) Both (A) and (R) are true and (R) is correct explanation of (A).
- (B) Both (A) and (R) are true and (R) is not correct explanation of (A).
- (C) (A) is true and (R) is false.
- (D) (R) is true and (A) is false.
- (iii) Which of the following is not a focus area of data governance ? 1
- (A) Data integrity (B) Data availability
- (C) Usability of data (D) Data visualisation



- (iv) In the K-Nearest Neighbors (KNN) algorithm, what does the “K” represent ? 1
- (A) K is the number of training examples.
(B) K is the number of points that the K-NN algorithm checks before doing a majority vote.
(C) K is the number of points that the K-NN algorithm checks before doing a minority vote.
(D) K is the dimensionality of the data.
- (v) Regression methods that do not have a linear function are called : 1
- (A) Linear regression models
(B) Non-linear regression models
(C) Exponential regression models
(D) Polynomial regression models
- (vi) What is the primary characteristic of Multiple Linear Regression ? 1
- (A) It uses only one independent variable to predict the outcome of a dependent variable.
(B) It uses multiple dependent variables to predict the outcome of a single independent variable.
(C) It uses multiple independent variables to predict the outcome of a dependent variable.
(D) It employs categorical variables exclusively for prediction.
3. Answer any 5 out of the given 6 questions. $5 \times 1 = 5$
- (i) True or False : A company is collecting and storing personal identifiable information in a secure, encrypted format without obtaining explicit consent from the user. There is no agreement from the user regarding the collection of the data by the company. It is a clear violation of data privacy rules. 1
- (ii) Statistical methods like mean, mode, median, range, variance and standard deviation are used for : 1
- (A) Bivariate Analysis (B) Univariate Analysis
(C) Multivariate Analysis (D) Exponential Analysis



- (iii) When using a decision tree for classification, what does each leaf node represent ? 1
- (A) A subset of the dataset (B) A class label
(C) A threshold value (D) A continuous variable
- (iv) What is the curse of dimensionality in the context of the K-Nearest Neighbors (KNN) algorithm ? 1
- (A) It refers to the phenomenon where KNN works well with a large number of input variables.
(B) It describes the ease with which KNN handles high-dimensional data.
(C) It suggests that KNN is not affected by the number of input variables.
(D) It explains that KNN works well with a small number of input variables but struggles as the number of variables grows.
- (v) The diagram below shows a scatter plot of two variables X and Y. Linear regression helps us to find Line X. What is Line X known as ? 1



- (A) Correlation line (B) Line of best fit
(C) Regression line (D) Trend line
- (vi) What problem-solving approach does the K-Means clustering algorithm use ? 1
- (A) Gradient Descent
(B) Expectation-Maximisation
(C) Principal Component Analysis (PCA)
(D) Genetic Algorithm



4. Answer any 5 out of the given 6 questions. $5 \times 1 = 5$
- (i) The European Union made _____ effective on May 25th, 2018 to protect European Union consumer data. 1
- (A) Health Insurance Portability and Accountability Act
(B) General Data Protection Regulation
(C) Personal Data Protection Bill
(D) California Consumer Privacy Act
- (ii) In a market survey we are analyzing the relationship between price and sales of a product to see if there is any relationship. This kind of analysis is a/an : 1
- (A) Descriptive Analysis (B) Univariate Analysis
(C) Multivariate Analysis (D) Bivariate Analysis
- (iii) What is the value of the terminal nodes in a regression tree after training ? 1
- (A) The minimum of the observations
(B) The maximum of the observations
(C) The mode of the observations
(D) The mean of the observations
- (iv) Why is the K-Nearest Neighbors (KNN) algorithm very sensitive to outliers ? 1
- (A) KNN relies on a fixed value of K, which makes it prone to outliers.
(B) KNN does not consider the distances between data points.
(C) KNN chooses neighbors based on distance criteria, and outliers can significantly affect distance calculations.
(D) KNN only works with categorical data, and outliers are more common in continuous data.
- (v) In real-life scenarios, why is it typically best to have a small Root Mean Squared Error (RMSE) value in regression models ? 1
- (A) A small RMSE indicates a higher Mean Absolute Error.
(B) A small RMSE indicates a stronger linear relationship between dependent and independent variables.
(C) A small RMSE signifies higher accuracy of the model's predictions.
(D) A small RMSE suggests that the model needs to be retrained.



- (vi) State True or False. 1

Clustering techniques apply when instances of data are to be divided into natural groups.

5. Answer any 5 out of the given 6 questions. 5 × 1 = 5

- (i) **Statement 1** : At any point of time, an individual can request an organization to remove all the data that they have collected about him/her. 1

Statement 2 : Data Privacy is just about secure data storage.

- (A) Both statement 1 and statement 2 are correct.
(B) Both statement 1 and statement 2 are incorrect.
(C) Statement 1 is correct and statement 2 is incorrect.
(D) Statement 2 is correct and statement 1 is incorrect.
- (ii) Identify the correct match. 1

- | | |
|--------------------------|-------------------------------|
| a. Supervised Learning | (i) Uses non-classified data |
| b. Unsupervised Learning | (ii) Clustering method |
| c. K-Means | (iii) Uses well labelled data |
- (A) a – (ii), b – (iii), c – (i)
(B) a – (i), b – (ii), c – (iii)
(C) a – (iii), b – (ii), c – (i)
(D) a – (ii), b – (i), c – (iii)

- (iii) Which of the following is NOT an advantage of K-NN algorithm ? 1

- (A) No training step (B) Sensitive to outliers
(C) Good for multi class problems (D) No assumptions of data

- (iv) Consider the following equation for Simple Linear Regression : 1

$$Y = m * X + b$$

Which of the following is correct with respect to the given equation ?

- (A) Y is an independent variable and X is a dependent variable.
(B) Y is an intercept and X is a dependent variable.
(C) X is an intercept and Y is a dependent variable.
(D) Y is a dependent variable and X is an independent variable.



- (v) Which of the following is not a real-world application of Unsupervised Learning ? 1
- (A) Image segmentation for medical diagnosis
(B) Recommendation systems
(C) Natural language translation
(D) Anomaly detection in network security
- (vi) The below mentioned formula represents : 1
- $$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \dots + \beta_p X_{ip} + \epsilon_i$$
- Where $i = 1, 2, 3, \dots, n$ for each observation.
- (A) Linear Regression (B) Non-Linear Regression
(C) Correlation (D) Multiple Linear Regression

Section – B

(Subjective Type Questions)

Answer any 3 out of the given 5 questions on Employability Skills. $3 \times 2 = 6$

6. Define the following terms : 2
- (A) Social Entrepreneur
(B) First generation entrepreneurs
7. Identify the type of sentence used in the statement as declarative, interrogative, exclamatory or imperative : 2
- (A) Complete your work
(B) It's very yummy !
(C) Did you complete your homework ?
(D) I received the prize for the best player in football team.
8. Why should a person adopt a positive attitude ? Give any two reasons. 2



9. Following table shows marks of 5 students of class IX in 3 subjects. Based on this data, calculate the following by writing a formula. 2

	A	B	C	D
	Name	Social Studies	Sanskrit	English
1.	Akash	79	91	89
2.	Avinash	75	95	81
3.	Rohit	69	80	70
4.	Mudit	82	90	71
5.	Yug	80	75	85

- (A) Total marks of Rohit and Mudit in Sanskrit.
(B) Total marks of Avinash in 3 subjects.
10. Define Solid waste management. Mention any two specific areas within waste management where green jobs are prevalent. 2

Answer any 4 out of the given 6 questions in 20-30 words each. 4 × 2 = 8

11. Explain the importance of data privacy in today's digital age and provide an example of a situation where a breach of data privacy could have significant consequences. 2
12. What are the two common techniques to handle missing data ? 2
13. List any 2 advantages of using a decision tree. 2
14. Name and explain the two important characteristics of KNN. 2
15. Describe the purpose of the Mean Absolute Error (MAE) in Linear Regression. 2
16. Name any four clustering methods. 2



Answer any **3** out of the given **5** questions in **50-80** words each. $3 \times 4 = 12$

17. Expand and explain COPPA. 4

18. Explain Univariate Analysis in detail. Give at least one example to support your answer. 4

19. Define a decision tree in the context of machine learning. Also, provide a labelled diagram of a simple decision tree. 4

20. Salma started a business 10 months back. It is observed that her income has increased exponentially every month. 4
 - (A) Which type of regression can you relate this to ?
 - (B) Write the formula for part (A) and explain the terms used.
 - (C) What will be the shape of graph formed in this case ?

21. Explain the following : 4
 - (A) Unsupervised Learning
 - (B) K-means Clustering

