

MASTER OF SCIENCE DATA SCIENCE AND ANALYTICS (MSCDSA)

MSCDSA/ASSIGN/SEMESTER-II

ASSIGNMENTS

(July – 2026 & January – 2027)

**MCS-066, MCS-067, MCS-224, MCS-068, MCSL-069,
MCSL-070**



**SCHOOL OF COMPUTER AND INFORMATION SCIENCES
INDIRA GANDHI NATIONAL OPEN UNIVERSITY
MAIDAN GARHI, NEW DELHI – 110 068**

CONTENTS

Course Code	Assignment No.	Submission-Schedule		Page No.
		For July-December Session	For January-June Session	
MCS-066	MSCDSA(II)/066/Assign/26-27	31 st October, 2026	30 th April, 2027	3
MCS-067	MSCDSA(II)/067/Assign/26-27	31 st October, 2026	30 th April, 2027	7
MCS-224	MSCDSA(II)/224/Assign/26-27	31 st October, 2026	30 th April, 2027	10
MCS-068	MSCDSA(II)/068/Assign/26-27	31 st October, 2026	30 th April, 2027	13
MCSL-069	MSCDSA(II)/L-069/Assign/26-27	31 st October, 2026	30 th April, 2027	16
MCSL-070	MSCDSA(II)/L-070/Assign/26-27	31 st October, 2026	30 th April, 2027	17

Important Notes

1. Submit your assignments to the Coordinator of your Study Centre on or before the due date.
2. Assignment submission before due dates is compulsory to become eligible for appearing in corresponding Term End Examinations. For further details, please refer to MSCDSA Programme Guide.
3. To become eligible for appearing the Term End Practical Examination for the lab courses, it is essential to fulfill the minimum attendance requirements as well as submission of assignments (on or before the due date). For further details, please refer to the MSCDSA Programme Guide.
4. The viva voce is compulsory for the assignments. For any course, if a student submitted the assignment and not attended the viva-voce, then the assignment is treated as not successfully completed and would be marked as ZERO.

Course Code : MCS-068
Course Title : Predictive Data Analysis
Assignment Number : MSCDSA (II)/068/Assign/2026-27
Maximum Marks : 100
Weightage : 30%
Last Dates for Submission : 31st October, 2026 (for July session)
 30th April, 2027 (for January session)

Note: There are twelve questions in this assignment, amounting to 80 marks, remaining 20 marks are for viva-voce. You may use illustrations and diagrams to enhance the explanations. Please go through the guidelines regarding assignments given in the MSCDSA Programme Guide.

Make suitable assumptions if any

Q1: Find the Quartile Deviation for the following data. (5 Marks)

Class Interval	Frequency
0–10	3
10–20	5
20–30	7
30–40	9
40–50	4

Q2: Referring to the Data given below, and Using Kendall's Tau correlation, a data scientist wants to diagnose whether the ranking of project complexity is related to the ranking for time-to-completion of the completed projects, (5 Marks)

Project	Complexity Rank (X)	Completion Time Rank (Y)
A	1	2
B	2	1
C	3	4
D	4	3

Q3: Consider the following **monthly sales data (in units)** for a product: (5 Marks)

Month	Sales
Jan	120
Feb	130
Mar	125
Apr	140
May	150
Jun	160

Compute a **3-month simple moving average** and interpret the result.

Q4: A factory produces two products A and B.

(8 Marks)

- Profit from Product A = ₹40 per unit
- Profit from Product B = ₹30 per unit

Each unit requires resources as follows:

Resource	Product A	Product B	Available
Labor Hours	2	1	100
Raw Material	1	1	80

Determine how many units of each product should be produced to maximize profit.

Q5: Consider the following Matrix representation of 2 documents and their associated shingles. Determine the Jaccard Similarity of the documents? **(7 Marks)**

Shingle/Documents	Set 1	Set 2
1	0	0
2	1	1
3	1	1
4	1	1
5	1	0

Q6: Determine the Inter Quartile Range (IQR) and comment on the existence of Outliers in the grouped data given below: **(8 Marks)**

Class Interval	Frequency
0–10	5
10–20	9
20–30	14
30–40	8
40–50	4

Q7: A teacher wants to study the relationship between hours studied and marks obtained by students. Here Hours Studied is the independent variable (X) and Marks Obtained is the Dependent variable (Y). Determine the equation of Linear Regression, using the data given below: **(7 Marks)**

Student	Hours Studied (X)	Marks Obtained (Y)
1	1	52
2	2	55
3	3	61
4	4	66
5	5	70

Q8: Consider the daily sales data of a shop recorded over three consecutive days, where the sales values are $Y_1 = 100$, $Y_2 = 110$, and $Y_3 = 125$. Assume that the data follows an ARIMA (1,1,1) model, where the parameters of the model are given as $\phi_1 = 0.6$, $\theta_1 = 0.4$, and the constant term $c = 2$. It is also given

that the error term at time $t = 2$ is $\epsilon_2 = 3$. Using the general form of the ARIMA model, forecast the value of sales for the fourth day, i.e., determine Y_4 . Show all necessary steps involved in the mapping of the AR, I, and MA components from the general ARIMA equation to the working forecasting equation. **(10 Marks)**

Q9: Suppose we have the following training dataset representing students based on two features: study hours and attendance percentage. The target variable indicates whether the student passed (Yes) or failed (No).

Study Hours	Attendance (%)	Result
2	50	No
4	60	No
6	70	Yes
8	80	Yes

Now, suppose a new student has studied for 5 hours and has 65% attendance. Predict whether this student will pass or fail using KNN with $K = 3$. **(8 Marks)**

Q10: Apply the K-Means Clustering algorithm to group the following data points into 2 clusters. The given data points are: $A(1,1)$, $B(2,1)$, $C(4,3)$, $D(5,4)$. Assume the initial centroids are: $C_1 = (1,1)$, $C_2 = (5,4)$. Use the Euclidean distance measure, perform the clustering step by step until the centroids do not change. Also write the final clusters and interpret the result. **(7 Marks)**

Q11: Suppose a disease screening model gives the following confusion matrix for 100 people:

- True Positive (TP) = 32
- False Positive (FP) = 10
- False Negative (FN) = 8
- True Negative (TN) = 50

Compute Accuracy, Precision, Recall, Sensitivity, Specificity and F1-Score **(5 Marks)**

Q12: Explain the following Data Structures in R **(5 Marks)**

- Vectors
- Lists
- Matrices,
- Arrays
- Frames