

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-066
Course Title	:	Mathematical Foundations for Data Science-II
Assignment Number	:	MSCDSA (II)/066/Assign/2026-27
Maximum Marks	:	100
Weightage	:	30%
Last Date of Submission	:	31st October, 2026 (for July session) 30th April, 2027 (for January session)

There are four questions in this assignment, which carry 80 marks. Each question carries 20 marks. Rest 20 marks are for viva voce. You may use illustrations and diagrams to enhance the explanations, if necessary.

Q1: (Covers Block 1)

(a) A retail company wants to study customer purchasing behaviour in its stores. What would be the population for this study be? How will you sample the data and draw inferences? Also, differentiate between descriptive and inferential statistics with suitable examples. **(2 Marks)**

(b) A researcher collects information on the following variables from 200 university students: **(2 Marks)**

- Gender
- Academic programme
- Annual family income
- Student satisfaction level (Very Low to Very High)
- CGPA

Identify the appropriate measurement scale (nominal, ordinal, interval or ratio) for each variable. Also, explain the importance of selecting the correct measurement scale during statistical analysis.

(c) The following marks (out of 100) were obtained by 30 students: **(5 Marks)**

42, 55, 67, 72, 48, 63, 59, 74, 81, 65, 58, 69, 77, 84, 71, 62, 57, 66, 73, 60, 75, 79, 68, 64, 53, 70, 61, 56, 76, 82

Construct:

- a frequency distribution,
- relative frequency table,
- histogram,
- ogive, and
- stem-and-leaf plot.

(d) Using the data given in Question 3, calculate: **(5 Marks)**

- Mean
- Median
- First Quartile (Q1)
- Third Quartile (Q3)

Draw a box plot and comment on the presence of any outliers.

- (e) The monthly sales (in lakh rupees) of a company for twelve months are: (4 Marks)

35, 42, 38, 41, 45, 39, 47, 50, 43, 44, 46, 40

Calculate:

- Range
- Variance
- Standard deviation
- Coefficient of variation

Comment on the sales pattern.

- (f) Explain the concepts of skewness and kurtosis. A data scientist observes two datasets with the same mean and variance but different skewness and kurtosis. Explain how these measures help in understanding the nature of the distributions. Also, explain the characteristics of a normally distributed dataset. (2 Marks)

Q2: (Covers Block 2)

- (a) If the letters of the word “Probability” are arranged randomly, what is the probability that all the letter ‘b’ are together in the random arrangement. (2 Marks)

- (b) A factory produces bulbs using two machines: (3 Marks)

- Machine A manufactures **60%** of the bulbs, and **2%** of the bulbs it produces are defective.
- Machine B manufactures **40%** of the bulbs, and **5%** of the bulbs produced by it are defective.

A randomly selected bulb is found to be defective. Using Bayes' theorem, determine the probability that the defective bulb was manufactured by Machine B.

- (c) A random variable (X) denotes the number of defective items found in a shipment and has the following probability distribution: (2 Marks)

X	0	1	2	3
P(X)	0.25	0.40	0.25	0.10

Calculate:

- Expected value of finding the
- Variance

- (d) Explain the concepts of joint probability mass function and marginal probability function for two discrete random variables with the help of an example. (2 Marks)

- (e) A manufacturing process produces 2% defective products. A random sample of 25 products is selected. Perform the following tasks for this process: (2 Marks)

- (i) Calculate the probability of obtaining:

- exactly one defective product,
- at most two defective products.

(ii) State the assumptions under which the binomial distribution is applicable.

- (f) Explain the Bernoulli, Binomial, Poisson and Hypergeometric distributions. Compare their assumptions, parameters and applications. **(3 Marks)**
- (g) The lifetime (in hours) of a sensor follows a normal distribution with a mean of 800 hours and a standard deviation of 60 hours. Compute the following for this sensor: **(3 Marks)**
- Probability that a sensor lasts more than 850 hours.
 - Probability that its lifetime lies between 760 and 900 hours.
- (h) Explain the Uniform, Normal, Chi-square, t and F distributions. Highlight the relationships among these distributions. **(3 Marks)**

Q3: (Covers Block 3)

- (a) Explain the Central Limit Theorem and discuss its significance in statistical inference. Illustrate how the theorem helps in estimating population parameters when the population distribution is unknown. **(3 Marks)**
- (b) A population has a mean of 80 and a standard deviation of 18. Random samples of size 64 are repeatedly selected and reported. Compute the following for this: **(3 Marks)**
- Mean of the sampling distribution of the sample mean.
 - Standard error.
 - Probability that the sample mean exceeds 84.
- (c) A random sample of 64 observations has a sample mean of 48 and a sample standard deviation of 10. Find the following: **(4 Marks)**
- A 95% confidence interval for the population mean.
 - Distinguish between point estimation and interval estimation.
- (d) Explain the principles of Maximum Likelihood Estimation (MLE). Also, explain the characteristics of an estimator. **(3 Marks)**
- (e) A company claims that the average battery life of its rechargeable battery is 18 hours. A random sample of 40 batteries gives a mean life of 17.2 hours with a standard deviation of 2.5 hours. Using a 5% significance level, test the manufacturer's claim. **(4 Marks)**

Clearly state the following while answering:

- Null and alternative hypotheses,
 - Test statistic,
 - Conclusion.
- (f) Explain the concepts of significance level, Type-I error, Type-II error and p-value. Also, compare the z-test and t-test, and discuss the situations in which each test should be applied. **(3 Marks)**

Q4: (Covers Block 4)

- (a) Three teaching methods are tested on three different groups of students. The examination scores are given below: **(4 Marks)**

- Method A: 72, 75, 70, 73, 74
- Method B: 78, 81, 79, 82, 80
- Method C: 69, 68, 71, 70, 67

Perform a one-way ANOVA at the 5% significance level and interpret the results.

- (b) Explain the assumptions underlying one-way and two-way ANOVA. Also, explain the two-way ANOVA using a suitable example. **(3 Marks)**
- (c) The following data show customer preference for payment method across two age groups: **(3 Marks)**

Age Group	UPI	Card	Cash
Below 40	60	45	25
40 and Above	35	40	55

Using the chi-square test of independence, determine whether payment preference is independent of age group at the 5% significance level.

- (d) Explain the objectives and procedures of: **(4 Marks)**
- Chi-square goodness-of-fit test,
 - Test of independence,
 - Kolmogorov–Smirnov goodness-of-fit test.
- (e) Explain the concepts of local minima, global minima, gradient descent and gradient ascent. Compare gradient descent and stochastic gradient descent. Discuss their advantages and limitations in machine learning applications. **(3 Marks)**
- (f) What are pseudo-random numbers? Explain the methods of generating pseudo-random numbers. **(3 Marks)**

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-067
Course Title	:	Data Wrangling and Visualization
Assignment Number	:	MSCDSA(II)/067/Assign/2026-27
Maximum Marks	:	100
Weightage	:	30%
Last Dates for Submission	:	31st October, 2026 (for July session) 30th April, 2027 (for January session)

There are four questions in this assignment, which carry 80 marks. Each question carries 20 marks. Rest 20 marks are for viva voce. Algorithms, if asked, should be written in pseudocode that is closer to Python. You may use illustrations and diagrams to enhance the explanations, if necessary

Q1:

- (a) Explain data wrangling. Discuss the importance of data semantics, set-based profiling, summary statistics, and data profiling in preparing raw data for data analysis. Illustrate your answer with a suitable example. **(3 Marks)**
- (b) A retail company has collected customer transaction data containing the following attributes: **(3 Marks)**
- Customer ID
 - Age
 - Gender
 - City
 - Purchase Amount
 - Payment Mode

Describe how you would perform data profiling on this dataset. What are the summary statistics that you would perform on this data?

- (c) Make a CSV file that contains employee information. The file should have a few missing salary values, a few duplicate employee records, a few inconsistent department names, and some incorrect city names. **(3 Marks)**

Write a Python program with suitable comments using **pandas** that performs the following tasks after reading the file.

- identify missing values,
 - remove duplicate records,
 - replace inconsistent values,
 - fill missing salary values using an appropriate method, and
 - rename the column headings.
- (d) Explain different techniques used for handling missing data. Discuss situations where each technique is appropriate. **(3 Marks)**

- (e) A dataset contains the ages of 500 customers ranging from 18 to 80 years. Explain different ways to classify these customers into age groups using discretisation (binning). **(4 Marks)**
- (f) Write Python code to detect outliers using the interquartile range method. Also, explain the strategies for handling outliers in data with the help of suitable examples. **(4 Marks)**

Q2:

- (a) Two data files contain customer details and purchase details having the same format. These files may contain overlapped data. Write and explain the Python program to merge the two files. Make and state assumptions, if any **(3 Marks)**
- (b) Explain hierarchical indexing. Illustrate how hierarchical indexing can be used for reshaping datasets with suitable examples. **(3 Marks)**
- (c) A supermarket dataset contains the following fields: **(3 Marks)**
- Region
 - Product Category
 - Sales
 - Profit

Write a Python program to perform the following tasks:

- group data by Region,
 - calculate total sales,
 - calculate average profit.
- (d) Why is the need to create groups during data analysis? Explain with the help of an example. Explain the working of the **groupby()** function in pandas. Discuss grouping using dictionaries, functions and Series in Python with suitable examples. **(3 Marks)**
- (e) Make a synthetic dataset containing monthly sales of three different products for one year. Write a Python program using Matplotlib and/or Seaborn to generate a line plot and a scatter plot. Make suitable assumptions. **(4 Marks)**
- (f) Write a Python code to demonstrate the purpose of figures, axes, labels, legends, annotations and subplot arrangements with suitable examples. **(4 Marks)**

Q3:

- (a) What is a scatter plot matrix? Why is it needed? Write a Python program to make a scatter plot matrix. **(3 Marks)**
- (b) A weather department has recorded daily rainfall and temperature data for five years. Suggest suitable graphical techniques for visualising this dataset. Justify your choices with reference to time series, scatterplot matrices, and geographical maps. **(6 Marks)**

- (c) Using Python, create a dashboard consisting of multiple plots arranged in a single figure. Demonstrate how titles, legends, annotations and customised plot parameters improve the readability of the visualisation. **(3 Marks)**
- (d) Explain any four applications of graphs in data science with suitable examples. **(4 Marks)**
- (e) A social networking website wants to analyse the relationships among users. Describe how graph representations such as bipartite graphs, hierarchical trees and graph drawing techniques can be used for analysing the network. Illustrate your answer with suitable diagrams. **(4 Marks)**

Q4:

- (a) Explain different methods of visualising random forests. **(5 Marks)**
- (b) Explain the use of the bivariate histogram and kernel density estimation with the help of a suitable example. **(6 Marks)**
- (c) Explain the Trellis Paradigm. Why is it needed? **(3 Marks)**
- (d) Explain various regression visualisations with the help of an example. **(6 Marks)**

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

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-224
Course Title : Artificial Intelligence & Machine Learning
Assignment Number : MSCDSA (II)/224/Assign/2026-27
Maximum Marks : 100
Weightage : 30%
Last Date of Submission : 31st October, 2026 (for July session)
 30th April, 2027 (for January session)

This assignment has sixteen questions with a total marks 80, answer all questions. The 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 Programme Guide for the format of presentation.

Q1: What is learning? Define the following ways of learning: (5 Marks)
 (I) Rote Learning (II) Learning by Instruction (III) Learning by analogy (IV) Learning by Induction (V) Learning by deduction

Q2: Define is Artificial Intelligence? What are the applications of AI in the healthcare and agricultural domains? (4 Marks)

Q3: Find the minimum cost path for the 8-puzzle problem, where the start and goal state are given as follows: (4 Marks)

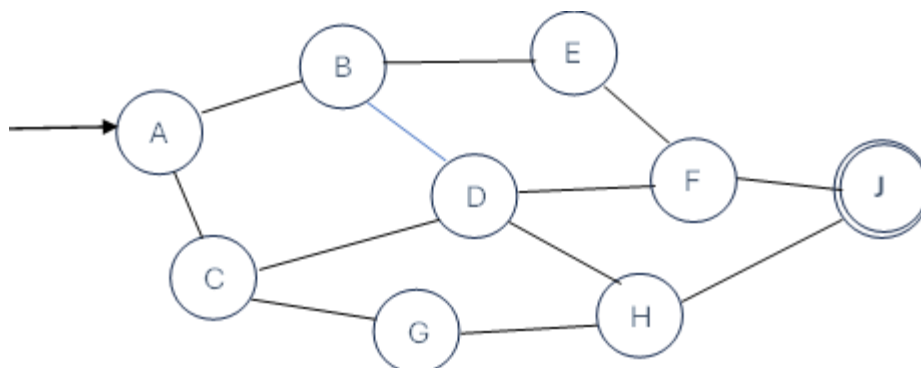
1	2	3
4	8	-
7	6	5

Start State

1	2	3
4	5	6
-	7	8

Goal State

Q4: Apply BFS algorithm on the following graph. (5 Marks)



Q5: Draw a semantic network for the following English statement: (4 Marks)
 "Shyam owns a dog named Sheru, and Sheru likes to chase cats".

Q6: In a class, three students tossed one coin (one each) 3 times. Answer the following. (4 Marks)

- (a) Write down all the possible outcomes which can be obtained in this experiment.
 (b) What is the probability of getting 2 or more than 2 heads at a time? Also, write the probability of getting three tails at a time.
 (c) Calculate the Relative frequency of tail $r_n(T)$.

Q7: Explain Dempster-Shafer theory with a suitable example. **(4 Marks)**

Q8: For the following fuzzy sets: **(5 Marks)**

$$A = \{ a/0.6, b/0.4, c/0.5, d/0.0, e/0.8 \}; B = \{ a/0.2, b/0.8, c/0.7, d/0.3, e/0.5 \}$$

$$C = \{ a/0.1, b/0.2, c/0.8, d/0.6, e/0.2 \}$$

Find the fuzzy sets: (i) $A \cup B \cup C$ (ii) $A \cap B \cap C$ (iii) $A' \cup B' \cup C'$ (iv) $A' \cap B' \cap C'$ (v) $(A \cap B \cup C)'$

Q9: What is ensemble learning? Explain three primary classes of ensemble learning methods. **(4 Marks)**

Q10: Use Naive Bayes classification method for the following dataset and classify the class (Species) of $X = \{ \text{Color=Green, Legs=2, Height=Tall, Smelly=No} \}$ **(7 Marks)**

Sl. No.	Color	Legs	Height	Smelly	Species
1	White	3	Short	Yes	M
2	Green	2	Tall	No	M
3	Green	3	Short	Yes	M
4	White	3	Short	Yes	M
5	Green	2	Short	No	H
6	White	2	Tall	No	H
7	White	2	Tall	No	H
8	White	2	Short	Yes	H

Q11: What is a Decision Tree? Use ID3 algorithm to create the decision tree for the following dataset and use it to find the class of unknown sample $X = \{ \text{Peter, red, short, average} \}$ **(7 Marks)**

Independent Attributes / Condition Attributes					Dependent Attributes / Decision Attributes
Name	Hair	Height	Weight	Lotion	Result
Sarah	blonde	average	light	no	sunburned (positive)
Dana	blonde	tall	average	yes	none (negative)
Alex	brown	short	average	yes	none

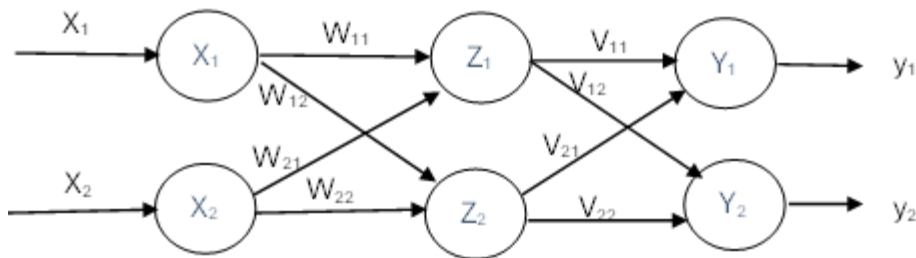
Annie	blonde	short	average	no	sunburned
Emily	red	average	heavy	no	sunburned
Pete	brown	tall	heavy	no	none
John	brown	average	heavy	no	none
Katie	blonde	short	light	yes	none

Q12: Find a quadratic regression model for the following data. Use the regression model and calculate the value of Y at X = 9. **(5 Marks)**

X	3	4	5	6	7
Y	2.5	3.2	3.8	6.5	11.5

Q13: For the given points of two classes Blue and Yellow: **(6 Marks)**
Blue: { (1,2), (2,3), (-1,2), (-1,4), (-1,-1) }
Yellow: { (4,2), (5,-1), (5,1), (6,1), (5,3) }
Plot a graph for the Blue and Yellow categories. Find the support vectors and optimal separating line.

Q14: The following diagram represents a feed-forward neural network with one hidden layer: **(6 Marks)**



For the following input patterns, calculate the output of the network (y_1 and y_2) if weights are initialized as: $W_{11} = -1$, $W_{12} = 2$, $W_{21} = 1$, $W_{22} = -2$, $V_{11} = 2$, $V_{12} = -1$, $V_{21} = -2$, $V_{22} = 2$. Use activation function $F(Y) = 1$ for $Y \geq 0$ and $F(Y) = 0$ otherwise.

Pattern	P ₁	P ₂	P ₃	P ₄
X ₁	0	1	0	1
X ₂	0	0	1	1

Q15: Consider the following two-dimensional pattern. Using PCA algorithm, calculate the principal component. **(5 Marks)**

X _i	2	4	5	6	6	7	9	8
Y _i	1	6	4	5	7	7	10	9

Q16: Explain Apriori algorithm with the help of a suitable example. **(5 Marks)**

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	:	MCSL-069
Course Title	:	Artificial Intelligence & Machine Learning Lab
Assignment Number	:	MSCDSA(II)/L-069/Lab_Assign/2026-27
Maximum Marks	:	100
Weightage	:	30%
Last Date of Submission	:	31st October, 2026 (for July session) 30th April, 2027 (for January session)

This assignment has 8 Questions. Answer all the questions. The total marks for all the questions are 40, and the maximum marks for each question are mentioned. Your Lab Records will carry 40 Marks. The 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 programme guide for the format of presentation.

Note: You must execute the program and submit the program logic, sample input, and output along with the necessary documentation. Assumptions can be made wherever necessary.

- Q1:** Write a Python Program to solve the N-Queen Problem without using Recursion. **(4 Marks)**
- Q2:** Implement the Water Jug problem in Python. **(4 Marks)**
- Q3:** Write a Python Program to implement the Min-Max Algorithm. **(4 Marks)**
- Q4:** Write a Python Program to implement the AO* Algorithm **(6 Marks)**
- Q5:** Discuss the Naïve Bayes algorithm and write the Python code to demonstrate the execution of the Naïve Bayes algorithm on the dataset of your choice. **(6 Marks)**
- Q6:** Write a Python Program to implement Logistic Regression for data classification, and choose a dataset of your own choice. **(4 Marks)**
- Q7:** Take a real-time example to implement the ID3 decision tree classification algorithm in Python. **(6 Marks)**
- Q8:** Write a Python Program to implement Support Vector Machines for data classification , and choose a dataset of your own choice. **(6 Marks)**

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 : **MCSL-070**
Course Title : **Data Analysis Lab**
Assignment Number : **MSCDSA(II)/L-070/Lab_Assign/2026-27**
Maximum Marks : **100**
Weightage : **30%**
Last Date of Submission : **31st October, 2026 (for July session)**
30th April, 2027 (for January session)

Note: This assignment consists of two sections, each carrying 20 marks, making a total of 40 marks. Your Lab Records will carry 40 Marks. The remaining 20 marks are for viva voce. Prepare and present your assignment in accordance with the guidelines provided in the MSCDSA Programme Guide. For programming-based questions, submit the source code, documentation, output, and screenshots of the program execution. You may also use diagrams, tables, illustrations, and screenshots, wherever appropriate, to improve the clarity and presentation of your answers.

SECTION 1: Data Wrangling and Visualisation Lab

1. Write Python programs using NumPy, Pandas, Matplotlib and/or Seaborn wherever required.
2. You may use Google Colab or Jupyter Notebook or any Python IDE.
3. Make and state assumptions, if any.

Q1:

(3 Marks)

A company that does retail marketing has collected customer transaction data, as shown in the following table

CustomerID	Name	City	Age	Amount	Payment Mode
C101	Amit	Delhi	25	1250	UPI
C102	Neha	Mumbai	32	2300	Card
C103	Ravi	Delhi	29	NaN	Cash
C104	Priya	Bengaluru	NaN	1850	Card
C105	Amit	Delhi	25	1250	UPI
C106	Karan	Delhi	41	4200	Cash
C107	Sonia	Kolkata	35	2750	UPI
C108	Meena	Delhi	27	1500	Card
C108	Meena	Delhi	27	1500	Card
C109	Rohan	Mumbai	31	3150	UPI

Perform the following tasks for the data as above:

- a) Create a DataFrame using the above data.
- b) Display summary statistics of the dataset.
- c) Find missing values and show records with missing data.

- d) Fill the missing numerical data using the mean value of that column.
- e) Remove all the duplicate records.
- f) Finally, display the cleaned dataset.

Q2:

(3 Marks)

The following synthetic sales data is for supermarket sales.

OrderID	Region	Category	Sales	Profit
O101	South	Grocery	5200	650
O102	South	Grocery	6100	720
O103	North	Clothing	18000	3100
O104	North	Grocery	4500	480
O105	North	Clothing	24000	5200
O106	South	Clothing	8200	1100
O107	South	Grocery	3900	420
O108	North	Clothing	7600	980
O109	North	Clothing	9100	1250
O110	South	Grocery	21000	4700

Perform the following tasks for the data as above:

- a) Group the data by Region and calculate the total Sales.
- b) What is the average Profit for each Category?
- c) Apply multiple aggregation functions (sum, mean, maximum) on the Sales column.
- d) Sort the aggregated data in descending order of Sales.
- e) Display the final aggregated table.

Q3:

(4 Marks)

Consider the following two datasets:

Dataset A: Student Information

StudentID	Name	Department
S101	Asha	Data Science
S102	Vivek	AI
S103	Mohit	Data Science
S104	Sneha	Cyber Security
S105	Riya	AI

Dataset B: Semester Marks

StudentID	Marks in Python	Marks in Statistics
S101	82	78
S102	74	80
S103	91	88
S104	69	72
S105	85	81

Perform the following tasks for the synthetic data as above:

- Merge the two datasets on StudentID.
- Create a pivot table showing department-wise average marks.
- Set Department and StudentID as a hierarchical index.
- Rearrange the index levels.
- Display the final reshaped DataFrame.

Q4:

(5 Marks)

The monthly sales (in thousands Rs) of three products are given below:

Month	Product A	Product B	Product C
Jan	45	52	39
Feb	48	55	41
Mar	52	58	45
Apr	49	56	47
May	60	63	51
Jun	64	66	54
Jul	67	70	58
Aug	65	72	60
Sep	70	75	63
Oct	74	78	67
Nov	78	82	71
Dec	81	86	74

Make the following plots by using Matplotlib and Seaborn libraries:

- Line plot
- Grouped bar chart
- Histogram for Product A sales
- Scatter plot between Product A and Product B
- Add suitable titles, legends, axis labels and annotations.

Q5:

(5 Marks)

The following synthetic dataset shows the data of advertisement expenditure and monthly sales:

Advertisement cost in Thousand Rs	Sales in Rs (lakhs)
5	21
8	24
10	27
12	30
14	33
16	35
18	38
20	41
22	43
24	46
26	48
28	51

Perform the following tasks for the synthetic data as above:

- Create a scatter plot of Advertisement versus Sales.
- Fit a simple linear regression model and draw the regression line.
- Plot the residuals.
- Plot the distribution of the Sales variable using the Kernel Density Estimate (KDE)

SECTION 2:

Q1:

(5 Marks)

A car showroom wants to analyse the details of different cars using R programming. The company wants to store car-related information, identify the type of each variable, convert variables into suitable data types, filter records, and summarise the data for decision-making.

Create or import the following dataset in R as car_showroom.csv:

Car_ID	Brand	Model	Mileage	Engine_Type	Transmission	Price_Lakh
C101	Maruti	Swift	22.5	Petrol	Manual	7.2
C102	Hyundai	i20	20.1	Petrol	Automatic	9.5
C103	Tata	Nexon	17.8	Diesel	Manual	11.3
C104	Honda	City	18.4	Petrol	Automatic	14.8

Car_ID	Brand	Model	Mileage	Engine_Type	Transmission	Price_Lakh
C105	Mahindra	XUV300	16.9	Diesel	Manual	12.6
C106	Toyota	Glanza	21.2	Petrol	Manual	8.9

Perform the following tasks in R:

- Import the dataset and display its structure using suitable R functions.
- Identify the data type and class of each variable using `typeof()`, `class()`, and `str()`.
- Convert Brand, Engine_Type, and Transmission into factor variables.
- Filter and display all cars having mileage greater than **18 km/litre**.
- Find the average mileage and average price of the cars.
- Convert all brand names into uppercase using suitable string functions.
- Create a new column named Price_Category with values **Low**, **Medium**, and **High** based on price.
- Write a short interpretation of how basic R data types and variables help in business data analysis.

Q2:

(5 Marks)

An e-commerce company wants to integrate customer, order, and product-review data stored in different files and visualise customer behaviour. The company wants to understand region-wise sales, category-wise product performance, and customer satisfaction.

Create the following three datasets in R:

Dataset 1: customers.csv

Customer_ID	Customer_Name	Region	Membership
CU101	Alisha	North	Gold
CU102	Ravi	West	Silver
CU103	Neha	South	Gold
CU104	Salim	East	Bronze
CU105	Priya	North	Silver

Dataset 2: orders.csv

Order_ID	Customer_ID	Product_Category	Amount
O1001	CU101	Electronics	45000
O1002	CU102	Clothing	8000
O1003	CU103	Grocery	3500
O1004	CU104	Electronics	52000
O1005	CU105	Clothing	12000
O1006	CU101	Grocery	5000

Dataset 3: reviews.csv

Review_ID	Customer_ID	Rating	Feedback
R01	CU101	5	Excellent
R02	CU102	3	Average
R03	CU103	4	Good
R04	CU104	5	Excellent
R05	CU105	2	Poor

Perform the following tasks in R:

- Import all three datasets using suitable R functions.
- Merge customer and order data using Customer_ID.
- Merge the resulting dataset with review data.
- Display the final integrated dataset.
- Find total sales region-wise and product-category-wise.
- Draw a bar chart showing total sales by region.
- Draw a pie chart or bar chart showing product-category-wise sales.
- Draw a scatter plot between Amount and Rating.
- Add suitable title, axis labels, colour, and legend wherever required.
- Interpret which region and product category performed best.

Q3:

(5 Marks)

An automobile manufacturing company wants to study vehicle safety, fuel efficiency, and transmission type using R. The company wants to apply statistical tests and predictive modelling to support product planning.

Use the built-in R datasets **Cars93** from the MASS package and **mtcars**. Perform the following tasks:

- Load the MASS package and import the Cars93 dataset.
- Extract the variables Type and Airbags.
- Create a contingency table between car type and airbag availability.
- Apply the Chi-Square test using `chisq.test()` and interpret whether car type and airbag availability are associated.
- Load the mtcars dataset and build a simple linear regression model using:
`lm(mpg ~ wt, data = mtcars)`
- Interpret how vehicle weight affects mileage.
- Predict mileage for cars having weights **2.5, 3.0, and 3.5**.
- Build a multiple regression model using:
`lm(mpg ~ disp + hp + wt, data = mtcars)`
- Identify which variable has the strongest effect on fuel efficiency.
- Build a logistic regression model to classify transmission type using:
`glm(am ~ hp + wt + cyl, data = mtcars, family = binomial)`
- Interpret which vehicle specifications influence manual or automatic transmission.
- Present all results with proper R output, graphs, and interpretation.