MASTER OF COMPUTER APPLICATIONS (MCA_NEW)

ASSIGNMENTS
OF MCA_NEW (2Yrs) PROGRAMME
SEMESTER-III

(January - 2025 & July - 2025)

MCS-224, MCS-225, MCS-226, MCS-227

MCSL-228, MCSL-229

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

CONTENTS

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 Programme Guide of MCA (2Yrs).
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 Programme Guide of MCA (2yrs).
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-226**
Course Title	**:**	**Data Science and Big Data**
Assignment Number	**:**	**MCA_NEW(III)/218/Assign/2025**
Maximum Marks	**:**	**100**
Weightage	**:**	**30%**
Last Dates for Submission	**:**	**30th April, 2025 (for January session)** **31st October, 2025 (for July session)**

This assignment has 10 questions of 8 Marks each, answer all questions. Rest 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: Define the term data science. Describe its applications in two industries of your choice (e.g., healthcare, finance, e-commerce). What role does the data science lifecycle play in managing data projects?

Q2: Explain Exploratory Data Analysis (EDA) and its importance. What are the main steps in performing EDA on a new dataset? Describe two methods for detecting outliers and how handling outliers impacts data analysis.

Q3: Describe the role of statistical hypothesis testing in data analysis. What are Type I and Type II errors, and how do they affect decision-making? Provide an example of hypothesis testing in a real-world scenario.

Q4: Discuss the 4 Vs of big data (Volume, Velocity, Variety, and Veracity). Provide a real-world example of each, explaining how these characteristics create challenges in big data management.

Q5: Explain the Hadoop architecture with a focus on HDFS and the master/slave architecture. How do NameNode and DataNodes work together to store and manage large datasets? Provide a basic example of this storage process.

Q6: Compare Apache Spark, Hive, and HBase in terms of functionality, data processing methods, and use cases. When would Spark be preferred over traditional MapReduce, and why?

Q7: Describe the purpose and functionality of a *Bloom filter* in data stream processing. How does the Bloom filter efficiently check for element presence? Describe the Flajolet-Martin algorithm for cardinality estimation in data streams.

Q8: What is the PageRank algorithm, and how is it used in link analysis? Describe the concept of "flow of rank" in PageRank. Explain how the PageRank of a web page is calculated using the flow model.

Q9: Discuss the challenges of online advertising and recommendation systems. Explain the concept of collaborative filtering with an example, and discuss the role of clustering in social network analysis.

Q10: What is the Random Forest algorithm? Explain how it can be applied to classification problems. Write a program in R to implement a Random Forest classifier on a sample dataset and explain its output.