

Punyashlok Ahilyadevi Holkar Solapur University, Solapur



Name of the Faculty: Science & Technology

(As per New Education Policy 2020)

Syllabus: - Data Science

Name of the Course: B.Sc. I (Sem. I &II)

(To be implemented from Academic Year 2024)

**Semester-wise Structure for
B. Sc. Data Science (Honors/Research) Programme
as per NEP-2020
(w.e.f. – June 2024)**

B.Sc. Part-I (Semester-I) Data Science						
Course Type	Course Code	Course Title	Credits	Teaching hours/week		
				T	P	Total
DSC	DSC-1	Introduction to DBMS G04-0113	2	2	--	2
	DSC-1 (P)	Data Science Practical-I G04-0113P	2	--	4	4
SEC	SEC-1	Introduction to R Programming G04-SEC-0113	2	--	4	4
GE/OE	GE/OE-1	Computer Organization and Operating Systems	2	2	--	2
B.Sc. Part-I (Semester-II) Data Science						
DSC	DSC-2	Introduction to Data Science G04-0213	2	2	--	2
	DSC-2 (P)	Data Science Practical-II G04-0213P	2	--	4	4
SEC	SEC-2	Introduction to Python Programming	2	--	4	4
GE/OE	GE/OE-2	Fundamentals of Data and Data Collection Techniques	2	2	--	2

B. Sc. Part-I (Data Science) Semester-I

DSC-1	Theory	INTRODUCTION TO DATA BASE MANAGEMENT SYSTEM	Credits: 02 Hours: 30
--------------	---------------	--	----------------------------------

Course Objectives:

- This course focuses on describing a sound introduction to the discipline of database management systems,
- This course gives a good formal foundation on the relational model of data and usage of Relational Algebra.

Course Outcomes:

After successful completion of this course, students are expected to:

- Explain the features of database management systems and Relational database.
- Create and populate a RDBMS for a real-life application, with constraints and keys, using SQL.
- Retrieve any type of information from a data base by formulating complex queries in SQL.
- Analyze the existing design of a database schema and apply concepts of normalization to design an optimal database.
- Build indexing mechanisms for efficient retrieval of information from a database.

Course Content

Unit-1: Basics of DBMS (15 hrs.)

Characteristics of database approach, data models, database system architecture and data independence. Entity types, relationships, constraints. Relational model concepts, relational constraints, relational algebra, SQL queries Mapping ER/EER model to relational database, functional dependencies, Lossless decomposition, Normal forms (upto BCNF).

Unit-2: File organization in Databases (15 hrs.)

ACID properties, concurrency control Operations on files, File of Unordered and ordered records, overview of File organizations, indexing structures for files (Primary index, secondary index, clustering index), Multilevel indexing using B and B+ trees.

Reference Books:

1. R. Elmasri, S.B. Navathe, *Fundamentals of Database Systems* 6 Edition, Pearson Education, 2010.
2. R. Ramakrishanan, J. Gehrke, *Database Management Systems* 3 Edition, McGraw-Hill, 2002.
3. A. Silberschatz, H.F. Korth, S. Sudarshan, *Database System Concepts* 6 Hill, 2010. Edition, McGraw-Hill
4. R. Elmasri, S.B. Navathe *Database Systems Models, Languages, Design and application Programming*, 6th Edition, Pearson Education, 2013.

DSC-1 (P)	Practical	DATA SCIENCE PRACTICAL-I	Credits: 02 Hours: 60
----------------------	------------------	---------------------------------	----------------------------------

Course Outcomes:

At the end of this course students are expected to be able to

1. Understand the framework of database management system.
2. Perform different operations on databases.

At least 10 practicals to be conducted based on DSC-1

SEC-1	Practical	Introduction to R Programming	Credits: 02 Hours: 60
--------------	------------------	--------------------------------------	----------------------------------

Course Objectives: The main objective of this course is to introduce the R software to the students. The course is aimed to inculcate analytical skills in students in order for them to fulfill current industry standards. At the end of the course, students are expected to be able,

1. To identify object types in R.
2. To input and edit data in R.
3. To format data in R.
4. To analyze data statistically in R.

Course Outcomes:

At the end of the course, students are able to

1. Use R in to solve real-world problems.
2. Create graphical representation of the data in using R.
3. Handle data using existing R functions.
4. Perform statistical analysis of the data using R.

Course Content List of Practical's

1	Simple manipulations (Numbers and Vectors, Vector algebra)
2	Matrices (Matrix algebra, Determinant, Inverse of matrix, Eigenvalues and Eigenvectors)
3	R packages and Data import-export (Loading and installing packages in R, Importing and exporting MS-Excel files)
4	Diagrammatic and Graphical Representation-I by using R (Frequency polygon, Histogram and Stem and leaf diagram)
5	Diagrammatic and Graphical Representation-II by using R (Bar and Pie diagrams)
6	Diagrammatic and Graphical Representation-III by using R (Line charts and Boxplots)
7	Repetitive execution in R: for loops, repeat and while

8	Descriptive statistics-I in R (AM, GM, HM, Mode, Median, Quartiles, Percentiles)
9	Descriptive statistics-II in R (Range, Standard deviation, Variance, Coefficient of variance)
10	Correlation and Regression

Reference Books:

1. Gardener, M. (2017). Beginning R: The statistical programming language, WILEY
2. Lawrence, M., & Verzani, J. (2016). Programming Graphical User Interfaces in R. CRC press. (ebook).
3. W. N. Venables, D. M. Smith and the R Core Team. (2023). An Introduction to R. (ebook)

GE/OE-1	Theory	Computer Organization and Operating Systems	Credits: 02
			Hours: 30

Course Objectives: The aim of this course is to provide learners with a foundational understanding of computer systems, covering both hardware organization and operating system fundamentals:

1. Understand the functional units of a computer, instruction execution, and data representation.
2. Explain the structure and organization of CPU, memory, and I/O subsystems.
3. Gain insights into the role of operating systems in process, memory, file, and device management.
4. Explore mechanisms for process scheduling, synchronization, and deadlock handling.
5. Investigate memory management techniques including paging, segmentation, and virtual memory.

Course Outcomes:

After completing the course, students will be able to:

1. Describe the structure, components, and functioning of a computer system.
2. Demonstrate knowledge of data representation, CPU design, memory hierarchy, and I/O subsystems.
3. Explain the role of the operating system in managing processes, memory, files, and devices.
4. Apply scheduling, synchronization, and deadlock handling algorithms to real scenarios.
5. Evaluate memory allocation, paging, segmentation, and replacement policies in OS.
6. Compare and analyze different disk scheduling techniques.

Course Content**Unit I – Computer Organization****[10 Hours]**

Introduction to Computer Organization: Basic functional units, von Neumann architecture, instruction cycle, instruction set classification (RISC vs CISC).

Data Representation: Number systems, binary arithmetic, complements, fixed-point and floating-point representation.

CPU Organization: Register organization, instruction formats, addressing modes, ALU design, control unit (hardwired vs microprogrammed).

Memory Organization: Primary, cache, and secondary memory, memory hierarchy, associative memory, cache mapping techniques, performance.

Input/Output Organization: I/O techniques – Programmed I/O, Interrupt-driven I/O, Direct Memory Access (DMA), I/O processors, buses.

Unit II – Operating Systems**[20 Hours]**

Introduction to Operating Systems: Definition, functions, types (Batch, Multiprogramming, Time-sharing, Real-time, Distributed, Parallel), OS services, system calls.

Process Management: Concept of process, process states, PCB, context switching, threads (types and benefits), operations on processes.

CPU Scheduling: Scheduling criteria, algorithms – FCFS, SJF, Round Robin, Priority Scheduling,

Multilevel Queue, Multilevel Feedback Queue.

Process Synchronization & Deadlocks: Critical section problem, semaphores, producer-consumer problem, classical problems (Reader-Writer, Dining Philosophers), deadlock characterization, prevention, avoidance (Banker's algorithm), detection & recovery.

Memory Management: Logical vs physical address space, dynamic loading, swapping, contiguous allocation, fragmentation, paging, segmentation, virtual memory, demand paging, page replacement (OPT, FIFO, LRU), thrashing.

Disk Management: Disk structure, disk scheduling algorithms (FCFS, SSTF, SCAN, C-SCAN).

Reference Books

1. Dhananjay M. Dhamdhere – *Systems Programming and Operating Systems*, McGraw Hill Education.
2. Andrew S. Tanenbaum – *Modern Operating Systems*, Pearson.
3. Abraham Silberschatz, Peter Galvin, Greg Gagne – *Operating System Concepts*, Wiley.
4. Carl Hamacher, Zvonko Vranesic, Safwat Zaky – *Computer Organization and Embedded Systems*, McGraw Hill.
5. M. Morris Mano – *Computer System Architecture*, Pearson.

B. Sc. Part-I (Data Science) Semester-II

DSC-2	Theory	INTRODUCTION TO DATA SCIENCE	Credits: 02 Hours: 30
--------------	---------------	-------------------------------------	--

Course Objectives:

- Introduction to Data Science and Analytics gives students a solid knowledge of the ideas, tools, and applications of data science across several domains.
- The course covers core concepts such data collection, preparation, analysis, visualization, and interpretation.
- Through hands-on activities and real-world case studies, students will obtain practical expertise with data analytics tools and programming languages typically used in data science.
- This course is designed to help students who want to advance their education or pursue professions in data science, machine learning, or artificial intelligence.

Course Outcomes:

Upon completion of Introduction to Data Science and Analytics, students will:

- Understand the fundamental principles and techniques of data science and analytics.
- Acquire proficiency in data acquisition, preprocessing, analysis, and visualization.
- Apply statistical methods and hypothesis testing techniques to analyze data and derive insights.
- Gain exposure to machine learning algorithms and their applications in solving real-world problems.
- Explore career opportunities in data science, machine learning, and artificial intelligence across various industries.

Course Content

Unit-1: Overview of data science and its applications. (15 hrs.)

- Role of data scientists and data analysts
- Introduction to data lifecycle and data-driven decision-making

Data Acquisition and Preprocessing

- Techniques for data collection and storage
- Data cleaning, transformation, and normalization
- Introduction to data wrangling tools (e.g., Pandas, SQL)

Unit-2: Exploratory Data Analysis (EDA) (15 hrs.)

Descriptive statistics and data visualization techniques

- Exploratory analysis using statistical plots and charts
- Hands-on exercises with data visualization.

Reference Books:

1. *"Data Science for Business"* book by Foster Provost and Tom Fawcett
2. *"Introduction to Statistical Learning"* book by Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani

3. *"Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow"* book by Aurélien Géron
4. *"Data Science from Scratch"* book by Joel Grus

DSC-2 (P)	Practical	DATA SCIENCE PRACTICAL-II	Credits: 02 Hours: 60
----------------------	------------------	----------------------------------	----------------------------------

Course Outcomes:

At the end of this course students are expected to be able to

1. Apply basic concepts of data science to real life problems.
2. Store and extract information from databases.

At least 10 practicals to be conducted based on DSC-2

GE/OE-2	Theory	Fundamentals of Data and Data Collection Techniques	Credits: 02 Hours: 30
----------------	---------------	--	----------------------------------

Course Objectives: The course aims to:

1. Provide foundational understanding of data, its types, and importance in decision-making.
2. Introduce methods of collecting primary and secondary data with their advantages and limitations.
3. Familiarize students with modern tools and technologies for data collection.
4. Explain the principles and methods of probability and non-probability sampling.
5. Develop awareness of issues related to data quality, reliability, validity, and bias.
6. Inculcate ethical practices in data collection and sampling through real-world case studies.

Course Outcomes:

After completing this course, students will be able to:

1. Identify and differentiate between types of data and their measurement levels.
2. Apply appropriate primary and secondary data collection methods for different contexts.
3. Utilize modern digital tools (Google Forms, SurveyMonkey, APIs) for efficient data collection.
4. Select and implement suitable probability and non-probability sampling techniques.
5. Evaluate data for reliability, validity, and bias, ensuring accuracy in decision-making.
6. Demonstrate ethical responsibility and apply data collection concepts to case studies from business, social sciences, and IT.

Course Content

Unit 1: Introduction to Data and Data Collections

(20 Hours)

Meaning and importance of data, Types of data: Primary vs Secondary, Qualitative vs Quantitative, Levels of data measurement: Nominal, Ordinal, Interval, Ratio, Role of data in decision-making and problem-solving

Data Collection Methods – Primary Data- Observation Method: participant vs non-participant, controlled vs uncontrolled, Survey Method: questionnaire design, interviews (structured, semi-structured, unstructured), Experimental Method: controlled experiments, field experiments, case studies, Advantages & limitations of primary data

Secondary Data-Sources of secondary data: government reports, company records, books, journals, websites, databases, Advantages and limitations of secondary data, Reliability, validity, and bias in secondary data, Big data sources (social media, IoT, web scraping basics)

Modern tools for data collection: Google Forms, SurveyMonkey, APIs, Trends: AI-based data collection, crowdsourcing

Unit 2: Sampling Techniques

(10 Hours)

Population vs Sample, Probability Sampling: simple random, stratified, cluster, systematic, Non-Probability Sampling: convenience, quota, snowball, judgmental, Sample size determination, errors in sampling, Ethical issues in data collection (privacy, informed consent, plagiarism), Case studies from business, social sciences, and IT

Reference Books: -

1. C.R. Kothari & Gaurav Garg – Research Methodology: Methods and Techniques, New Age International.
2. William G. Zikmund – Business Research Methods, Cengage.
3. Donald R. Cooper & Pamela S. Schindler – Business Research Methods, McGraw Hill.
4. Navin K. Sharma – Data Collection Methods and Research Methodology.
5. Alan Bryman & Emma Bell – Business Research Methods, Oxford University Press.
6. S. P. Gupta – Statistical Methods, Sultan Chand & Sons.

SEC-2	Practical	INTRODUCTION TO PYTHON PROGRAMMING	Credits: 02 Hours: 60
-------	-----------	---	----------------------------------

Course Objectives:

1. To get acquainted with Python programming.
2. To understand objects and structures in python.

Course Outcomes: Upon successful completion of this course, a student will be able to:

- i) Develop programmes in Python.
- ii) Perform data organization, data manipulation, statistical and mathematical computations, and data analysis using Python.

Course Content List of Practical's

1	Create NumPy arrays from Python Data Structures, Intrinsic NumPy objects and Random Functions.
2	Manipulation of NumPy arrays- Indexing, Slicing, Reshaping, Joining and Splitting
3	Import a CSV file and perform various Statistical and Comparison operations on rows/columns.

4	Create Pandas Series and Data Frame from various inputs
5	Import any CSV file to Pandas Data Frame and perform the various operations.
6	Visualize data using Line Plots, Bar Plots, Histograms, Density Plots and Scatter Plots.
7	Visualize data using Pie charts, Boxplots and Hitmap.
8	Descriptive statistics-I in Python (AM, GM, HM, Mode, Median, Quartiles, Percentiles)
9	Descriptive statistics-II in python (Range, Standard deviation, Variance, Coefficient of variance)
10	Correlation and Regression

Reference Books:

1. Gowrishankar S. & Veena, A. (2018). Introduction to Python programming. CRC Press.
2. Guttag J. V. (2021). Introduction to Computation and Programming Using Python: With Application to Computational Modeling and Understanding Data. Mit Press.
3. Haslwanter T. (2016). An Introduction to Statistics with Python. With Applications in the Life Sciences. Switzerland: Springer International Publishing.
4. Nelli F. (2018). Python data analytics with Pandas, NumPy, and Matplotlib.
5. Unpingco J. (2016). Python for probability, statistics, and machine learning (Vol. 1). Springer International Publishing.
6. Vander Plas, J. (2016). Python data science handbook: Essential tools for working with data. " O'Reilly Media, Inc."