MINOR
Subject: BSc- Data Science (2023-24)
COURSE STRUCTURE

Year	Semester	Paper	Subject	IA	EA	Practical	Total
1	**II**	**1A**	Principles of Data Science	25	75	-	100
2	**III**	**III A**	Fundamental of Statistics	25	75	-	150
	IV	**IV A**	Mathematics For Data Science	25	75	-	150
		IV B	Introduction to Data Science With R	25	75	50	150
			Introduction to Data Science With R Practical course				
3	**V**	**V A**	Statistical Methods	25	75	-	100
		V B	Big Data technology	25	75	50	150
			Big Data technology Practical course				

B.Sc -Data Science

Syllabus for Semester II

MINOR
B.Sc Data Science – I Year
II Semester Paper: I A

PRINCIPLES OF DATA SCIENCE

COURSE OBJECTIVES:

To provide strong foundation for data science and application area related to information technology and understand the underlying core concepts and emerging technologies in data science

COURSE OUTCOMES:

Upon completion of this course, the students should be able to:

1. Explore the fundamental concepts of data science
2. Understand data analysis techniques for applications handling large data
3. Understand various machine learning algorithms used in data science process
4. Visualize and present the inference using various tools
5. Learn to think through the ethics surrounding privacy, data sharing and algorithmic decision-making

UNIT-1-INTRODUCTION TO DATA SCIENCE (9HRS)

Definition – Define Data Science and Introduction to Data Science. - Data Science Process Overview – Defining goals – Retrieving data – Data preparation – Data exploration – Data modeling – Presentation.

UNIT-2 -BIG DATA (9HRS)

Problems when handling large data – General techniques for handling large data – Steps in big data – Distributing data storage .

UNIT-3-MACHINE LEARNING (9HRS)

Machine learning – Modeling Process – Training model – Validating model – Predicting new observations .

UNIT-4-DEEP LEARNING (9HRS)

Introduction – Deep Feedforward Networks – Regularization – Optimization of Deep Learning – Applications of Deep Learning.

UNIT-5 - DATA VISUALIZATION (9HRS)

Introduction to data visualization – Data visualization options – Filters – MapReduce – Dashboard development tools .

TEXT BOOKS:

1. Introducing Data Science, Davy Cielen, Arno D. B. Meysman, Mohamed Ali, Manning Publications Co., 1st edition, 2016
2. An Introduction to Statistical Learning: with Applications in R, Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, Springer, 1st edition, 2013
3. Deep Learning, Ian Goodfellow, Yoshua Bengio, Aaron Courville, MIT Press, 1st edition, 2016
4. Ethics and Data Science, D J Patil, Hilary Mason, Mike Loukides, O' Reilly, 1st edition, 2018

REFERRENCE BOOKS:

1. Data Science from Scratch: First Principles with Python, Joel Grus, O'Reilly, 1st edition, 2015
2. Doing Data Science, Straight Talk from the Frontline, Cathy O'Neil, Rachel Schutt, O' Reilly, 1st edition, 2013.
3. Mining of Massive Datasets, Jure Leskovec, Anand Rajaraman, Jeffrey David Ullman, Cambridge University Press, 2nd edition, 2014

B.Sc -Data Science

Syllabus for Semester III

MINOR
B.Sc Data Science – II Year
III Semester Paper: III A

FUNDAMENTALS OF STATISTICS

COURSE OBJECTIVES:
To enable the students to understand the fundamentals of statistics to apply descriptive measures and probability for data analysis.

COURSE OUTCOMES:
Upon completion of this course, the students should be able to:

1. Understand the science of studying & analyzing numbers.
2. Identify and use various visualization tools for representing data.
3. Describe various statistical formulas.
4. Compute various statistical measures.

UNIT I - Statistics: (12HRS)
Introduction to Statistics – Origin of Statistics, Features of Statistics, Scope of Statistics, Functions of Statics, Uses and importance of Statistics.

UNIT II - Collection of Data: (12HRS)
Introduction to Collection of Data, Primary and Secondary Data, Methods of Collecting Primary Data, Methods of Secondary Data.

UNIT III - Classification of Data Frequency Distribution :(12HRS)
Introduction Classification of Data, Objectives of Classification, Methods of Classification. Diagrammatic and Graphical Presentation of Data: Introduction to Tabular Presentation of Data, Objectives of Tabulation, Table, Types of Tables, Introduction to Diagrammatic Presentation of Data, Advantage and Disadvantage of Diagrammatic Presentation.

UNIT IV - Measures of Central tendency: (12HRS)
Introduction to Central Tendency, Purpose and Functions of Average, Types of Averages, Meaning of Arithmetic Mean,Median,Mode,Geometric Mean, Harmonic Mean- Properties Merit and Demerits and Simple Problems.

UNIT V - Measures of Dispersion: (12HRS)
Meaning of Dispersion, Range, Mean Deviation, Standard Deviation, Quartile Deviation- Properties Merit and Demerits and Simple Problems.

TEXT BOOKS:

1. Statistics and Data Analysis, A.Abebe, J. Daniels, J.W.Mckean, December 2000.
2. Statistics, Tmt. S. EzhilarasiThiru, 2005, Government of Tamilnadu.
3. Introduction to Statistics, David M. Lane.
4. Weiss, N.A., Introductory Statistics. Addison Wesley, 1999.
5. Clarke, G.M. & Cooke, D., A Basic course in Statistics. Arnold, 1998.

REFERENCE BOOKS:

1. Banfield J.(1999), Rweb: Web-based Statistical Analysis, Journal of Statistical Software.
2. Bhattacharya,G.K. and Johnson, R.A.(19977), Statistical Concepts and Methods, New York, John Wiley & Sons.

B.Sc -Data Science

Syllabus for Semester IV

MINOR

B.Sc Data Science – II Year

IV Semester Paper: IV A

MATHEMATICS FOR DATA SCIENCE

COURSE OBJECTIVES:

- To understand the difference between various types of matrices.
- To learn the basic concept and applications of matrices in real life problems.
- To identify and practice the mathematical logic problems with the help of truth tables or without using truth tables.
- Ability to implement features of inference rules in inference calculus.
- To understand the concept of Boolean algebra and Boolean functions.
- To understand the concepts of graphs, directed graphs, and trees.

COURSE OUTCOME:

At the end of the course student will be able to

- Determine inverse of matrix, perform matrix operations, solving systems of simultaneous linear equations (L3)
- Demonstrate concepts of mathematical logic for analyzing propositions and proving theorems. (L2)
- Analyze logical propositions via truth tables. (L3)
- Understand the basic properties of Boolean algebra and simplify simple Boolean functions using the basic Boolean properties. (L2)
- Understand and apply the fundamental concepts in graph theory in solving practical problems. (L3)
- Model problems in Computer Science using graphs and trees. (L3)

Syllabus:

UNIT - I

Matrices: Definition, addition and multiplication of matrices, various types of matrices, Determinant of a square matrix, Inverse of a matrix, Solution of system of non-homogenous linear equations by Crammer's rule , matrix inversion method, Gauss-Jordan method.

UNIT - II

Mathematical Logic: Connectives, Negation, Conjunction, Disjunction, Conditional &Bi-Conditional, Well Formed Formulae, Tautologies, Equivalence of formulae, Duality, Tautological Implications, Functionally Complete Set of Connectives.

UNIT - III

Boolean algebra: Definition and Examples, sub algebra, Direct product and Homomorphism, Boolean Functions, Boolean forms and free Boolean Algebras, Values of Boolean expressions

and Boolean functions, Representation of Boolean functions, Minimization of Boolean functions, Karnaugh maps. (8hours)

UNIT - IV

Graph Theory: Definitions, Finite and Infinite graphs, Incidence and Degree, Isolated pendant vertices, Isomorphism, sub graphs, Walk, Path and Circuit, Connected and Disconnected graphs, components, Euler graphs, Euler graph theorem.

UNIT-V

Trees: Properties of trees, pendant vertices, distance & centers, rooted & binary trees, spanning trees, fundamental circuit, shortest spanning trees, Kruskal's algorithm, Binary Tree Traversals. (8 hours)

Text Books :

1. Higher Engineering Mathematics by B.S.Grewal, Khanna Publishers, 43rd edition, 2015.
2. Numerical methods for scientific and engineering computation by M.K.Jain, S.R.K. Iyengar, R.K. Jain, New Age International publishers, 6thedition,2012.
3. Discrete Mathematical Structures with Applications to Computer Science by J.P. Tremblay and R. Manohar, Tata McGraw Hill,1997.
4. Graph Theory with Applications to Engineering and Computer Science by Narsingh Deo, Prentice Hall of India,2006.

Reference Books:

1. Discrete Mathematics and its Applications by Keneth. H. Rosen, Tata McGraw-Hill, 6th Edition, 2009.
2. Discrete Mathematics by Richard Johnsonbaug, Pearson Education, 7th Edition,2008.
3. Discrete Mathematics for Computer Scientists and Mathematicians by J.L. Mott, A.Kandel, T.P. Baker, PrenticeHall.

MINOR
B.Sc Data Science – II Year
IV Semester Paper: IV B

Introduction to Data science With R

Objective

Data Science is a fast-growing interdisciplinary field, focusing on the analysis of data to extract knowledge and insight. This course will introduce students to the collection. Preparation, analysis, modeling and visualization of data, covering both conceptual and practical issues. Examples and case studies from diverse fields will be presented, and hands-on use of statistical and data manipulation software will be included.

Outcomes

1. Recognize various disciplines that contribute to a successful data science effort.
2. Understand the processes of data science - identifying the problem to be solved, data collection, preparation, modeling, evaluation and visualization.
3. Be aware of the challenges that arise in data sciences.
4. Develop and appreciate various techniques for data modeling and mining.
6. Be cognizant of ethical issues in many data science tasks.
7. Be comfortable using commercial and open source tools such as the R language and its associated libraries for data analytics and visualization.
8. Learn skills to analyze real time problems using R
9. Able to use basic R data structures in loading, cleaning the data and preprocessing the data.
10. Able to do the exploratory data analysis on real time datasets
11. Able to understand and implement Linear Regression
12. Able to understand and use - lists, vectors, matrices, data frames, etc.

Syllabus:

Unit-1:
Introduction to Data Science- Introduction- Definition - Data Science in various fields - Examples - Impact of Data Science - Data Analytics Life Cycle - Data Science Toolkit - Data Scientist - Data Science Team

Unit-2:
Introduction to R- Features of R - Environment - R Studio. Basics of R-Assignment - Modes - Operators - special numbers - Logical values - Basic Functions - R help functions - R Data Structures - Control Structures.

Unit-3:

Matrices - Creating Matrices - Adding or removing rows/columns - Reshaping - Operations - Special functions on Matrices. Lists - Creating List – General List Operations - Special Functions - Recursive Lists.

Unit- 4:

Descriptive Statistics - Central Tendency - Variability - Mean - Median - Range - Variance - Summary - Handling Missing values and Outliers - Normalization

Data Visualization in R : Types of visualizations - packages for visualizations - Basic Visualizations, Advanced Visualizations and Creating 3D plots.

Unit- 5:

Inferential Statistics with R - Types of Learning - Linear Regression- Simple Linear Regression - Implementation in R - functions on lm() - predict() - plotting and fitting regression line. Multiple Linear Regression - Introduction - comparison with simple linear regression - Implementation of Multiple Linear Regression in R.

References

1. Nina Zumel, John Mount, “Practical Data Science with R”, Manning Publications, 2014.
2. .Jure Leskovec, Anand Rajaraman, Jeffrey D.Ullman, “Mining of Massive Datasets”, Cambridge University Press, 2014.
3. 3.Mark Gardener, “Beginning R - The Statistical Programming Language”, John Wiley & Sons, Inc., 2012.
4. W. N. Venables, D. M. Smith and the R Core Team, “An Introduction to R”, 2013. 5.Tony Ojeda, Sean Patrick Murphy, Benjamin Bengfort, Abhijit Dasgupta, “Practical Data Science Cookbook”, Packt Publishing Ltd., 2014.
5. Nathan Yau, “Visualize This: The FlowingData Guide to Design, Visualization, and Statistics”, Wiley, 2011.
6. Boris lublinsky, Kevin t. Smith, Alexey Yakubovich, “Professional Hadoop Solutions”.

Introduction to Data science With R – Practices

R Programming LAB

1) Installing R and R studio

2) Create a folder DS_R and make it a working directory. Display the current working directory

3) Installing the "ggplot2", "caTools", "CART" packages

	C1	C2	C3	C4	C5
C1	0	12	13	8	20
C2	12	0	15	28	88
C3	13	15	0	6	9
C4	8	28	6	0	33
C5	20	88	9	33	0

- Find the pairs of cities with shortest distance.

4) Load the packages "ggplot2", "caTools".

5) Basic operations in r

6) Working with Vectors:

- Create a vector v1 with elements 1 to 20.
- Add 2 to every element of the vector v1.
- Divide every element in v1 by 5
- Create a vector v2 with elements from 21 to 30. Now add v1 to v2.

7) Getting data into R, Basic data manipulation

8) Using the data present in the table given below, create a Matrix "M"

Find the pairs of cities with shortest distance.

9) Consider the following marks scored by the 6 students

Section	Student no	M1	M2	M3
A	1	45	54	45
A	2	34	55	55
A	3	56	66	64
B	1	43	44	45
B	2	67	76	78
B	3	76	68	37

- Create a data structure for the above data and store in proper positions with proper names
- Display the marks and totals for all students
- Display the highest total marks in each section.
- Add a new subject and fill it with marks for 2 sections.

Three people denoted by P1, P2, P3 intend to buy some rolls, buns, cakes and bread. Each of them needs these commodities in differing amounts and can buy them in two shops S1, S2. The individual prices and desired quantities of the commodities are given in the following table "demand.

	price	
	S1	S2
Roll	1.5	1
Bun	2	2.5
Cake	5	4.5
Bread	16	17

	demand.quantity			
	Roll	Bun	Cake	Bread
P1	6	5	3	1
P2	3	6	2	2
P3	3	4	3	1

- Create matrices for above information with row names and col names.
- Display the demand. Quantity and price matrices
- Find the total amount to be spent by each person for their requirements in each shop
- Suggest a shop for each person to buy the products which is minimal.

10) Consider the following employee details:

employee details as follows		
	emp_no:1	
	name: Ram	
	salary	
		basic: 10000
		hra: 2500
		da: 4000
	deductions	
		pf: 1100
		tax: 200
	total salary	
		gs(Gross Salary):
		ns(Net Salary)

- Create a list for the employee data and fill gross and net salary.
- Add the address to the above list
- display the employee name and address
- remove street from address
- remove address from the List.

B.Sc -Data Science

Syllabus for Semester V

MINOR
B.Sc Data Science – III Year
V Semester Paper: V A

STATISTICAL METHODS

COURSE OBJECTIVES:

At the end of the course, the students will be able to:

- Knowledge of Statistics and its implementation through practical understanding for various domains related to data science.
- Knowledge of various types of data, their organization and evaluation of summary measures such as measures of central tendency and dispersion etc.
- Knowledge of other types of data reflecting quality characteristics including concepts of independence and association between two attributes, insights into preliminary exploration of different types of data.
- Knowledge of correlation, regression analysis, regression diagnostics, partial and multiple correlations.

Syllabus:

UNIT –I : Curve fitting:

Bi- variate data, Principle of least squares, fitting of degree polynomial. Fitting of straight line, Fitting of Second degree polynomial or parabola, Fitting of power-curve and exponential curves.

UNIT-II: Correlation:

Meaning, Types of Correlation, Measures of Correlation: Scatter diagram, Karl Pearson's Coefficient of Correlation, Rank Correlation Coefficient (with and without ties) and Simple Problems.

UNIT III: Regression:

Concept of Regression, Linear Regression: Regression lines, Regression coefficients and it's properties, Regressions lines for bi-variate data and simple problems.

UNIT-IV: Attributes:

Notations, Class, Order of class frequencies, Ultimate class frequencies, Consistency of data, Conditions for consistency of data for 2 and 3 attributes only , Independence of attributes.

UNIT-V: Attributes:

Association of attributes and its measures, Relationship between association and colligation of attributes, Contingency table: Square contingency, Mean square contingency, Coefficient of mean square contingency**.**

Text book and Reference books:

1. V.K. Kapoor and S.C.Gupta: Fundamentals of Mathematical Statistics, Sultan Chand &Sons, New Delhi.
2. BA/B.Sc I year statistics-descriptive statistics, probability distribution-Telugu Academy-Dr M.Jaganmohan Rao, Dr N.Srinivasa Rao, Dr P.Tirupathi Rao, Smt.D.Vijayalakshmi.
3. K.V.S.Sarma: Statistics

REFERENCEBOOKS:

1. WillamFeller: Introduction to Probability theory and its applications. Volume–I,Wiley
2. Goon AM, GuptaMK, Das GuptaB: Fundamentals of Statistics,Vol-I,the World Press Pvt.Ltd., Kolakota.
3. HoelP.G: Introduction to mathematical statistics, Asia Publishing house.
4. M.Jagan Mohan Rao and PapaRao: ATextbook of Statistics Paper-I.
5. Sanjay Arora and Bansi Lal: New Mathematical Statistics: Satya Prakashan, NewDelhi

MINOR
B.Sc Data Science – III Year
V Semester Paper: VB

BIG DATA TECHNOLOGY

COURSE OBJECTIVES:

This course provides practical foundation level training that enables immediate and effective participation in big data projects. The course provides grounding in basic and advanced methods to big data technology and tools, including MapReduce and Hadoop and its ecosystem.

COURSE OUTCOME:

- Learn tips and tricks for Big Data use cases and solutions.
- Acquire knowledge of HDFS components, Name node, Data node, etc.
- Acquire knowledge of storing and maintaining data in cluster, reading data from and writing data to Hadoop cluster.
- Able to maintain files in HDFS
- Able to write MapReduce applications to access data present on HDFS
- Able to read different formats of files into map-reduce application.
- Able to develop MapReduce applications to analyze Big Data related to the real world use cases.
- Able to write MapReduce applications that can take data from multiple datasets and join them
- Able to optimize the performance of Map-Reduce application

Syllabus:

UNIT – I: Introduction to Big Data

Introduction –Distributed File System – Big Data and its importance, Characteristics of Big Data, Limitation of Conventional Data Processing Approaches, Need of big data frameworks, Big data analytics, Limitations of Big Data and Challenges, Big data applications.

UNIT – II: Hadoop:

Basic Concepts of Hadoop and its features -The Hadoop Distributed File System (HDFS)- Anatomy of a Hadoop Cluster - Hadoop cluster modes - Hadoop Architecture, Hadoop Storage - Hadoop daemons (Name node-Secondary name node-Job tracker-Task tracker-Data node,etc).

UNIT – III: Hadoop Ecosystem Components

Schedulers- Fair and Capacity, Hadoop 2.0 Vs Hadoop 3.0 and its new features.

Hadoop Cluster Setup – SSH & Hadoop Configuration –HDFS Administering – Monitoring & Maintenance.

UNIT – IV : Hadoop MapReduce

Introduction - Phases in MapReduce Framework - Anatomy of MapReduce Job run - Failures, Job Scheduling, Shuffle and Sort, Task Execution, Map Reduce ,Types and Formats, Map Reduce Features.

UNIT-V

Writing first MapReduce Program - Hadoop's Streaming API - Using Eclipse for Rapid Development – YARN Vs MapReduce Advanced MapReduce Concepts: Partitioner – Combiner – Joins – Map-side Join – Reduce-side Join.

Text Books :

References

1. Boris lublinsky, Kevin t. Smith Alexey Yakubovich, "Professional Hadoop Solutions". Wiley, ISBN : 9788126551071, 2015.
2. Chris Eaton, Dirk Deroos et al., "Understanding Big Data", McGraw Hill , 2010.
3. Tom White, "HADOOP" : The definitive Guide", O Reilly 2012.
4. Srinath Perera, Thilina Gunarathne, "Hadoop MapReduce Cookbook", PACKT publishing, 2013

BIG DATA TECHNOLOGY -LAB

1. **Case Study I**: Centers for Medicare & Medicaid Services: The Integrity of Healthcare Data and Secure Payment Processing.
2. **Case Study II**: Movie Lens Data set Analysis

RECOMMENDED CO-CURRICULAR ACTIVITIES:

A. Measurable

1. Assignments (in writing and doing forms on the aspects of syllabus content and outside the syllabus content. Shall be individual and challenging)
2. Student seminars (on topics of the syllabus and related aspects (individual activity))
3. Quiz (on topics where the content can be compiled by smaller aspects and data (Individuals or groups as teams))
4. Study projects (by very small groups of students on selected local real-time problems pertaining to syllabus or related areas. The individual participation and contribution of students shall be ensured (team activity

B. General

1. Group Discussion
2. Try to solve MCQ's available online.
3. Others

RECOMMENDED CONTINUOUS ASSESSMENT METHODS:

Some of the following suggested assessment methodologies could be adopted;

1. The oral and written examinations (Scheduled and surprise tests)
2. Closed-book and open-book tests
3. Problem-solving exercises
5. Practical assignments and laboratory reports
6. Observation of practical skills
7. Individual and group project reports like "Movie Lens Data Analysis", "Youtube Click stream Data Analysis", etc.
8. Efficient delivery using seminar presentations,
9. Viva voce interviews.
10. Computerized adaptive testing, literature surveys and evaluations.

Prepared by:

S. Rama Devi (PhD)

HOD, Department of Statistics