REGULATIONS AND CURRICULUM GOVERNING

MASTER OF SCIENCE PROGRAM

(Academic year from which the curriculum is implemented/ revised/ amended)

Name of the Constituent Unit/Centre: DEPARTMENT OF STATISTICS



YENEPOYA
(DEEMED TO BE UNIVERSITY)

Deralakatte, Mangalore, Karnataka.

NOTIFICATION FROM THE UNIVERSITY

YENEPOYA (DEEMED TO BE UNIVERSITY)

OFFICE OF THE REGISTRAR

No.: 55-ACM/07/2024

Date: 15.04.2024

NOTIFICATION

Sub: Starting of Post Graduate Program "M.Sc. Data Science" – reg .:-

Ref: 1. Resolution of the Academic Council at its 55th meeting held on 15.02.2024 vide agenda-7.
2. Resolution of the Executive Council at its 2nd meeting held on 13.03.2024 vide agenda 04.

*****

Pursuant to the resolutions cited in the above references, the Yenepoya (Deemed to be University) hereby notifies starting of Post Graduate Program "M.Sc. Data Science" in the department of Statistics under the Faculty of Science from the academic year 2024-2025.

REGISTRAR

mj

To:

PA to Vice Chancellor
Dean of Academics, Yenepoya (Deemed to be University)
Controller of Examinations
Dean, Faculty of Science
Professor & Head, department of Statistics
File Copy

YENEPOYA

YENEPOYA (DEEMED TO BE UNIVERSITY)

No.: 60-ACM/12/2024

OFFICE OF THE REGISTRAR
Date: 18.12.2024

NOTIFICATION

Sub: Regulations and Curriculum approval of M.Sc. Data Science program - reg .:-

Ref: 1. Resolution of the Academic Council at its 60th meeting held on 19.11.2024 vide agenda – 12

2. Resolution of the Executive Council at its 6th meeting held on 17.12.2024 vide agenda-05.

*****

Pursuant to the resolutions cited in the above references, the Yenepoya (Deemed to be University) hereby notifies the Regulations and Curriculum approval of M.Sc. Data Science program under the Faculty of Science from the academic year 2024-2025.

REGISTRAR

To:

PA to Vice Chancellor
Dean of Academics, Yenepoya (Deemed to be University)
Controller of Examinations
Dean, Faculty of Science
File Copy

YENEPOYA
(DEEMED TO BE UNIVERSITY)

CONTENTS

S. No.	Particulars	Page numbers
1.	Preamble	**1**
2.	Eligibility for Admission	**1**
3.	Aim of the Program	**1**
4.	Objectives/Outcomes of the program (Knowledge Skill/Attitude/ Communication)	**2-3**
5.	Program Structure	**4**
6.	Attendance	**5**
7.	Syllabus/Course Contents	**5-8**
8.	Scheme of Examination	**8-10**
9.	Eligibility to appear for University Examination	**11**
10.	Re-Entry after Break of the Study	**12**
11.	Maximum period for completion of the Program	**12**
12.	Criteria for Pass	**12**
13.	Letter Grades	**13**
14.	Academic Calendar	**14**
15.	Detailed Syllabus	**15-56**

YENEPOYA
(DEEMED TO BE UNIVERSITY)

1. Preamble

Data Science, an emerging and highly in-demand field, integrating advanced Statistical Methods, Machine Learning, Artificial Intelligence, and Computer Science to analyze large datasets, driving decision-making and innovation across industries. Professionals in Data Science are sought after in sectors such as healthcare, finance, marketing, cybersecurity, and e-commerce, where data-driven insights play a critical role in gaining a competitive edge. The M.Sc. in Data Science equips graduates for roles like Data Scientist, Data Analyst, Data Engineer, and Machine Learning Specialist, providing them with the robust technical and analytical skills valued by industries and organizations. To meet the growing demand for skilled data scientists, YENEPOYA (Deemed to be University) has launched a Post Graduate program in Data Science within the Department of Statistics, focused on delivering comprehensive training in statistical and computational methodologies essential for both academic and industry success.

2. Eligibility for Admission

A candidate who has passed an Undergraduate degree with 50 % aggregate marks from any University in India or abroad that is recognized by UGC/AIU. Students must fulfill either criteria A or B described below in order to be eligible for the program:

A. BE/B.Tech/B.Sc. degree with Computer Science/ Mathematics/ Statistics as major or minor (minimum of two years of learning)

OR

B. Bachelor of Computer Applications (BCA) / Bachelor degree in Data Science

3. Aim of the Program

The aim of the M.Sc. Data Science program is to equip students with essential skills and knowledge to analyze complex datasets and solve real-world problems. The program emphasizes a strong foundation in Statistics, Machine Learning, and Artificial Intelligence, alongside proficiency in modern data science tools. Students will develop critical thinking, ethical awareness, and effective communication, preparing them for successful, adaptable careers in data science.

YENEPOYA
(DEEMED TO BE UNIVERSITY)

4.Objectives/Outcomes of the program (Knowledge Skill/Attitude/ Communication)

- Program Outcomes

Code	Attribute	Program Outcome
PO1	Strong Theoretical Foundation (K)	Develop a strong theoretical foundation in statistics, data analysis, and computational techniques essential for data science.
PO2	Machine Learning and AI Expertise (K)	Acquire expertise in machine learning and artificial intelligence algorithms to solve complex data-related problems
PO3	Data Driven Decision Making (A)	Utilize data-driven decision making processes to address real-world challenges, demonstrating a commitment to data ethics and integrity.
PO4	Problem Solving and Critical Thinking (A)	Foster problem-solving and critical thinking skills to effectively analyze and interpret data, leading to informed decision-making.
PO5	Cutting Edge Technologies (S)	Gain proficiency in cutting-edge technologies and tools used in data science, including big data frameworks and cloud computing.
PO6	Effective Communication (C)	Communicate complex data insights and findings effectively to diverse audiences through presentations and reports.
PO7	Lifelong Learning and Adaptability (S)	Cultivate lifelong learning and adaptability to stay current with evolving data science methodologies and technologies.
PO8	Self-Directed Learning (S)	Engage in self-directed learning to independently acquire new skills and knowledge in data science.
PO9	Ethical Awareness (A)	Demonstrate ethical awareness in data handling and analysis, making responsible decisions in line with professional standards.
PO10	Leadership and Teamwork (A)	Exhibit leadership and teamwork abilities in collaborative projects, promoting a cooperative learning environment.

- Program Specific Outcomes

PSO1: Mastery of Statistical and Mathematical Foundations: Graduates will acquire a robust theoretical and practical understanding of core mathematical and statistical concepts essential for data science.

PSO2: Expertise in Machine Learning and Deep Learning: Graduates will demonstrate a strong command of data science methodologies. They will develop skills in machine learning, big data analytics, and deep learning to extract insights from complex datasets.

PSO3: Advanced Programming and Data Handling Skills: Graduates will attain proficiency in a variety of programming languages and tools essential for data sciences such as Python, R, SQL, mySQL, and NoSQL databases.

PSO4: Applied Research and Problem-Solving Abilities: Graduates will develop advanced research and problem-solving skills, utilizing key analytical techniques based on designs of experiments, regression analysis, multivariate analysis, Bayesian inference, time series analysis, categorical data analysis to address complex, data-driven challenges. These skills will enable graduates to design, conduct, and interpret research studies effectively.

PSO5: Multidisciplinary Application of Data Science Knowledge: Graduates will be equipped to apply data science solutions across interdisciplinary fields such as pharmacy, healthcare, finance, environmental science, and social sciences, utilizing techniques from statistics, machine learning, and deep learning to address complex challenges in these diverse domains.

YENEPOYA
(DEEMED TO BE UNIVERSITY)

5.Program Structure

The following table gives the program structure of M.Sc. Data Science

	Broad category of course	**Credits**	**Course (Title + Course Code)**
1.	**Discipline-Specific Courses**		
a)	Theory	**48**	Annexure I
b)	Practical (SEC)*	**18**	Annexure I
c)	Clinical (SEC)*	**-**	-
d)	Simulation-based (SEC)*	**-**	-
2.	**Curriculum Enrichment Courses**		
a)	Field Visit	**-**	-
b)	Community Posting	**-**	-
c)	Industry / Clinical Internship & Project work	**15**	-
d)	Research Project	**-**	-
3.	**Electives**		
a)	Discipline-Specific Electives	**09**	Annexure I
b)	Interdisciplinary Courses	**06**	Annexure I
c)	Ability Enhancement Courses	**-**	-
d)	Value Added Courses	**-**	-
Total		**96**	

6.Attendance

Each course (theory, practical, etc.) shall be treated as an independent unit for the purpose of attendance. A student shall attend a minimum of 80% of the total instruction hours in a course including tutorials and seminars in each semester.

7.Syllabus/Course Contents

Semester I

Course Code	Type of Course	Course Name	L	P	C	Marks		Total
						CIA	SEE	
Theory								
FS02DS-1C1	HC	Mathematical Foundation	4	0	4	50	50	100
FS02DS-1C2	HC	Statistical Foundation for Data Science	4	0	4	50	50	100
FS02DS-1C3	HC	Principles of Data Science	4	0	4	50	50	100
FS02DS-1C4	HC	Database Management Systems	4	0	4	50	50	100
Practical								
FS02DS-1P1	HC	Practical I: Programming Using Python and Data management using SQL	0	6	3	50	50	100
FS02DS-1P2	HC	Practical II: Data Analysis Using R	0	6	3	50	50	100
Total Credits								**22**
L: Lecture, P: Practical, C: Credits, HC: Hard Core, CIA: Continuous Internal Assessment, SEE: Semester End Examination								

Semester II

Course Code	Type of Course	Course Name	L	P	C	Marks		Total
						CIA	SEE	
Open Elective								
FS02DS-2E1	SC	Essential Statistics for Research	3	0	3	50	50	100
Theory								
FS02DS-2C1	HC	Multivariate Analysis	4	0	4	50	50	100
FS02DS-2C2	HC	Machine Learning	4	0	4	50	50	100
FS02DS-2C3	HC	Applied Regression Analysis	4	0	4	50	50	100
FS02DS-2C4	HC	Data warehousing and Data Mining	4	0	4	50	50	100
Practical								
FS02DS-2P1	HC	Practical III: Based on FS02DS-2C1 and FS02DS-2C3 using Python	0	6	3	50	50	100
FS02DS-2P2	HC	Practical IV: Based on FS02DS-2C2 and FS02DS-2C4 using Python	0	6	3	50	50	100
Total Credits								**25**
L: Lecture, P: Practical, C: Credits, HC: Hard Core, CIA: Continuous Internal Assessment, SEE: Semester End Examination								

Semester III

Course Code	Type of Course	Course Name	L	P	C	Marks		Tot
						CIA	SEE	
Open Elective								
FS02DS-3E1	SC	Applied Statistics for Research	3	0	3	50	50	100
Hard Core Theory								
FS02DS-3C1	HC	Time Series Analysis	4	0	4	50	50	100
FS02DS-3C2	HC	Deep Learning	4	0	4	50	50	100

FS02DS-3C3	HC	Big Data Analytics	4	0	4	50	50	100
Soft Core Theory- Elective I								
FS02DS-3C4.1	SC	Survival Analysis	3	0	3	50	50	100
FS02DS-3C4.2	SC	Non parametric regression	3	0	3	50	50	100
Soft Core Theory- Elective II								
FS02DS-3C5.1	SC	Design of Experiments and Clinical Trials	3	0	3	50	50	100
FS02DS-3C5.2	SC	Categorical Data Analysis and Bayesian Inference	3	0	3	50	50	100
Practical								
FS02DS-3P1	HC	Practical V: Based on Statistical theory papers using R	0	6	3	50	50	100
FS02DS-3P2	HC	Practical VI: Based on FS02DS-3C2	0	6	3	50	50	100
Total Credits								**27**
L: Lecture, P: Practical, C: Credits, HC: Hard Core, CIA: Continuous Internal Assessment, SEE: Semester End Examination								

Semester IV

Course Code	**Type of Course**	**Course Name**	**L**	**P**	**C**	**Marks**		**Tot**
						CIA	**SEE**	
Hard Core Theory								
FS02DS-4C1	HC	Research Methodology	4	0	4	50	50	100
Soft Core Theory- Elective III								
FS02DS-4C2.1	SC	Cloud Computing	3	0	3	50	50	100

FS02DS-4C2.2	SC	Internet of Things	3	0	3	50	50	100
FS02DS-4C2.3	SC	Applied Artificial Intelligence	3	0	3	50	50	100
Project Work								
FS02DS-4C3	HC	Project Work	0	0	15	150	150	300
Total Credits								**22**
L: Lecture, P: Practical, C: Credits, HC: Hard Core, CIA: Continuous Internal Assessment, SEE: Semester End Examination								

8.Scheme of Examination

The evaluation consists of two main components

A. Continuous Internal Assessment (CIA): The marks for internal assessment shall be awarded based on seminars, journal paper presentations, tests, assignments etc. The assessment gives importance to continuous and comprehensive evaluation. Each theory and practical course carry an internal assessment of 50 marks which is based on two sessional tests, seminars/ Journal paper presentation, Case studies/Data Analysis and assignments. The project work shall be assessed out of 150. The internal assessment marks shall be notified to the students before it is communicated to the Controller of Examinations, before the commencement of the University examinations. The split up is as follows:

Components of CIA	**Details**	**Weightage**
Theory		
Sessional Tests	Average of the two tests	25
Assignment	Course related topics	08
Seminar	Seminar / One Journal paper relevant to the core courses	08
Skill Enhancing Exercise	Case studies / Data Analysis /Outreach activities/other creative activities	09
Total		50

Practical		
Sessional Tests	Average of the two tests	25
Assignment	Practical related topics	10
Regularity	Punctuality in practical session and record submission	5
Performance	Performance in Practical session/ Internal viva	10
Total		50
Project Work		
Preliminary work	Topic Selection and Justification (10), Research Objectives and Scope (10), Initial Outline/Structure (5)	25
Literature Review	Range of Sources (10), Critical Analysis and Integration (15), Alignment with Research Goals (5)	30
Punctuality and Consistency	Meeting Milestones/Deadlines (10), Active Engagement and Updates (5)	15
Presentations	Two presentations (25+25)	50
Writing and documentation		30
Total		150

B. Semester End Examination (SEE) [Pattern-Theory, Practical and Project]: There shall be Semester End Examination (SEE) at the end of each semester ordinarily during December/January for odd semesters and during June/July for even semesters. The SEE duration shall be three hours. The University examination is for 50 marks of 3 hours' duration. The project work evaluated out of 150 marks. Pattern of question paper for semester end examination is as follows.

Key Criteria	**Details**	**Maximum Marks**
Theory		
Short answer type questions of 2 marks each (Answer 5 out of 7 questions)	5 × 2 marks	**10**

Long answer type questions of 10 marks each (Answer 4 out of 6 questions)	4 × 10 marks	**40**
Total		**50**
Practical		
Practical Record and Viva	-	**10**
Answer 4 out of 6 questions.	4 × 10	**40**
Total		**50**
Project Work		
Documentation/Thesis - Introduction/ Background (10) - Literature Review (10) - Materials and Methods (20) - Results (20) - Conclusion (15)		**75**
Presentation and Viva		**75**
Total		**150**

Valuation of Answer Scripts

a. Each theory examination shall be evaluated by one internal and one external examiner. There shall be a third evaluation if the difference is more than 15%.

b. Practical examination shall be jointly conducted and evaluated by one internal examiner and one external examiner.

Evaluation of Project

A) In-semester assessment is with a weight of 150 marks for the continuous assessment of candidate's performance regarding progress during project work, presentation and execution of project. At the end of the fourth semester the students need to submit two copies of their project report in the format prescribed

by the department.

B) End-semester examination with a weight of 150 marks is based on documentation, presentation and viva-voce shall be jointly conducted and evaluated by one internal examiner and one external examiner.

9. Eligibility to appear for University Examination

- Candidates having minimum 80% attendance in each of the courses can only qualify to appear for the semester end examinations. The candidates shall register for all the courses of a semester when he/she appears for the examination of that semester for the first time.
- Student who has less than 80% attendance shall not be permitted to appear for the Semester end examination in the course in which the short fall exists. The HOD/course coordinator through the Dean of Faculties shall announce the names of the students who will not be eligible to take the Semester End-Examinations in the various courses and send a copy of the same to the **m**Controller of Examination's (CEO) office. Registrations of such students for those courses shall be treated as cancelled. These students should repeat the course by attending classes in the succeeding years. Candidate who fails to meet minimum attendance of 80% in any of the Course /project work shall repeat that Course/project work subsequently.

10.Re-Entry after Break of the Study

i. Students admitted to a program abstaining for more than 3 months must seek readmission into the appropriate semester.

ii. The student shall follow the syllabus in vogue (currently approved /is being followed) for the program.

iii. All re admissions of students are subject to the approval of the University.

11.Maximum period for completion of the Program

A candidate shall complete the four semesters (two years) program within five years from the date of admission.

12.Criteria for Pass

i. A candidate shall be declared to have passed the PG program if he/she secures at least a CGPA of 5.0 (Course letter Grade P) in the aggregate of both internal assessment and semester end examination marks.

ii. For each course the total of 100% is determined from the CIA evaluation and the SEE. The minimum pass mark for CIA shall be 50% and SEE shall be 50%.

iii. The candidates who pass all the semester examinations in the first attempt in two years are eligible for ranks provided they secure at least a CGPA of 6.0 (at least letter Grade A).

iv. The results of the candidates who have passed the fourth semester examination but not passed the lower semester examinations shall be declared as NCL (Not Completed Lower semester examinations). Such candiates shall be eligible for the degree only after completion of all the lower semester examinations.

v. A candidate who passes the semester examinations in parts is eligible for only CGPA and letter Grade but not for ranking.

vi. Carry over shall be allowed to the candidates for all the semesters.

vii. Candidate who fails in any of the Course /project work shall reappear in that Course/project work and pass the examination subsequently.

13.Letter Grades

The results of successful candidates at the end of each semester shall be declared in terms of Grade Point Average (GPA) and alpha sign grade. The results at the end of the fourth semester shall be classified based on the Cumulative Grade Point Average (CGPA) obtained in all the four semesters and the corresponding overall alpha-sign grade. The letter grade as described below shall be adopted.

CGPA Range	Program alpha-sign grade
9.0-10.0	O+(Outstanding)
8.0 – 8.99	O(Excellent)
7.0 - 7.99	A+ (Very Good)
6.0 - 6.99	A(Good)
5.5 - 5.99	B^+ (Average)
5.0 – 5.49	P (Pass)
<5.0	F (Fail)

Calculation of Cumulative Grade Point Average (CGPA): It is a measure of overall cumulative performance of a student over all semester. The CGPA is the ratio of total credit points secured by a student in various courses in all semesters and the sum of all the total credits of all courses in all the semesters. It is expressed up to two decimals.

$$\frac{\sum_{i=1}^{4} C_i S_i}{\sum_{i=1}^{4} C_i}$$

where Si is the SGPA of ith Semester and C_i is the total number of credits in that semester

14. Academic Calendar

Semesters: An academic year shall consist of two semesters

Odd Semester 1st& 3rd	July/August to December/January
Even semester 2nd& 4th	January/February to June/July

Hours of instructions in classroom teaching, theory, practical, project work etc. per week

Semester	Theory	Practical	Project work	Total Hours	Total credits
I	16	12	-	28	22
II	19	12	-	31	25
III	21	12		33	27
IV	7	-	15	37	22
Total	**63**	**36**	**15**	**-**	**96**

Yenepoya (Deemed to be University) M.Sc. Data Science (2025) FS02DS-1C1: Mathematical Foundation		
Teaching Scheme: Theory: 4 hrs/week		**Credits:** 4

Course Outcomes: Students will be able to

CO1. Define and apply set operations and measure theory concepts.

CO2. Analyse properties of sets and functions using Lebesgue measure.

CO3. Assess convergence and divergence of sequences and series.

CO4. Perform operations on series and derive approximations.

CO5. Explain continuity and differentiability of functions.

CO6. Apply Riemann-Stieltjes integration and related techniques.

CO7. Understand vector spaces and orthogonal projections.

CO8. Execute matrix operations and solve linear equations.

CO9. Describe advanced concepts like characteristic roots and matrix decompositions.

CO10.Analyse and classify quadratic forms effectively.

Course Contents

Unit I	**Set Theory**	**(08Hours)**
Definition, basics of set operations, counting principles. Sequences of sets and their limits. Set function; Measure, Lebesgue measure, measurable functions.		
Unit II	**Numerical sequences and series**	**(10 Hours)**
Sequences and their convergence; sub sequences and Cauchy sequences. Series- convergence and divergence, tests for convergence, addition and multiplication of series.		
Unit III	**Advanced Calculus**	**(10 Hours)**
Continuity and differentiation of functions. Differentiation of vector valued functions. Special functions; Power series, series expansion and approximations. The Riemann Stieltjes‘ integration, properties. Integration by parts, change of variables, step functions as integrators.		

Unit IV	Linear Algebra and Matrix theory	(16 Hours)
Vector spaces, linear dependence and independence; basis and dimension of a vector space. Ortho normal basis and orthogonal projections. Gram-Schmidts Orthogonalization Process. Matrix Operations: Elementary Row operations; Row reduced Echelon Form, Rank and inverse of a matrix. Partitioned matrices; Generalized inverse of matrices. Linear equation-homogenous and non-homogenous systems, solution		
Unit V	**Advanced matrix theory**	**(12 Hours)**
Characteristic roots and vectors, Cayley-Hamilton theorem, Determinant, rank and trace of a matrix in terms of characteristic roots. Diagonalization; LU Decomposition; Cholesky Decomposition, Spectral Decomposition, Singular value decomposition. Positive definite matrices. Real quadratic forms, classification of quadratic forms, reduction of quadratic forms.		

References:

1. Apostol TM. Mathematical Analysis. N.Pl.: Pearson Education Asia and China Machine Press; 2004.
2. Bapat RB. Linear Algebra and Linear Models. Springer Science & Business Media; 2008.
3. Khuri AI. Advanced Calculus with Applications in Statistics. LibreDigital; 2003.
4. Lay D, Lay S, McDonald J. Linear Algebra and Its Applications. Pearson; 2014 May.
5. Malik SC, Arora S. Mathematical Analysis. New Age International; 1992.
6. Rao CR. Linear Statistical Inference and Its Applications. John Wiley & Sons, Inc. eBooks; 1973.
7. Searle SR. Matrix Algebra Useful for Statistics. Hoboken, New Jersey: Wiley; 2017.

Mapping of COs to POs

	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10
CO1	✓									
CO2	✓			✓				✓		
CO3	✓									

CO4	✓			✓				✓		
CO5	✓									
CO6	✓			✓				✓		
CO7	✓									
CO8	✓			✓				✓		
CO9	✓									
CO10	✓									

Yenepoya (Deemed to be University)

M.Sc. Data Science (2025)

FS02DS-1C2: Statistical Foundation for Data Science

Teaching Scheme:		**Credits:**
Theory: 4 hrs/week		4

Course Outcomes: Students will be able to

CO1. Apply the principles of probability to calculate the likelihood of events using classical, frequency, and axiomatic approaches.

CO2. Demonstrate understanding of random variables and their properties, including cumulative distribution, probability mass, and density functions, and compute conditional probabilities using Bayes' theorem.

CO3.Explain different discrete probability distributions and apply their properties to solve real-world problems.

CO4. Compute moments, moment-generating functions, and characteristic functions for discrete distributions to summarize and interpret data behaviour.

CO5.Examine the properties and applications of key continuous probability distributions including their moments and characteristic functions.

CO6. Utilize the Law of Large Numbers and the Central Limit Theorem to interpret sampling distributions and apply these concepts in statistical data analysis.

CO7. Distinguish between parametric and non-parametric estimation methods and evaluate the properties of point estimators, such as unbiasedness, consistency, and efficiency

CO8. Conduct and interpret parametric tests (e.g., t-tests, ANOVA) for continuous data analysis.

CO9. Construct confidence intervals for mean, variance, and proportions, analysing the implications for large and small sample sizes and non-normal data distributions.

CO10. Interpret confidence intervals and levels in hypothesis testing, and apply these concepts to compare means and proportions across different populations and scenarios.

Course Contents

Unit I	Elementary Probability Theory	(14Hours)
Random Experiment; sample point; sample space; events, frequency and classicaldefinitions of probability. Axiomatic definition of probability; addition and multiplicationtheorems; conditionalprobability and independence; Bayes' theorem (thrust is on numericalproblems &applications).Random variable, cumulative distribution function, probability mass function, probability density function. Joint distributions, marginal distributions, conditional distribution, expectation of a random variable and its properties, conditional expectation, addition and multiplication theorem of conditional expectation, moments and moment generating function, characteristics function, and probability generating function.		
Unit II	**Discrete Probability Distributions**	**(14 Hours)**
Discrete distributions: Binomial, Poisson distribution, Uniform distribution,Negative Binomialdistribution, Hyper geometric distribution, Geometric distribution andMultinomial distributions(Definition, moments, properties, moment generating function, Characteristic function and application).		
Unit III	**Continuous Probability Distributions**	**(12 Hours)**
Continuous distributions: Normal distribution, Student t distribution, F distribution, Chi- square distribution, Exponential distribution, Weibull distribution, Beta distribution, gamma distribution and lognormal distribution (Definition, moments, properties, characteristic function, and applications). Concept of non-central distributions Chi-square, t and F. Sampling distributions - The Concept and applications of Law of large numbers, central limit theorem (CLT) and applications.		
Unit IV	**Theory of Estimation**	**(10 Hours)**
Theoretical framework– parametric and non-parametric approach, Introduction to estimation, random sample, statistic. Point and interval estimation. Properties of good estimators. Sufficiency – factorization theorem. Likelihood function. Fisher information. Unbiasedness. Standard error of an estimator. Mean squared error. Uniformly Minimum Variance Unbiased Estimator (UMVUE). Cramer Rao lower bound. Minimum Variance Bound Unbiased Estimator (MVBUE). Efficiency. Consistency. Asymptotic efficiency. Asymptotic Relative Efficiency. Best Asymptotically Normal (BAN) estimator.		

Method of Methods. Maximum Likelihood Estimators (MLE). Properties of MLEs.		
Unit V	**Interval Estimation**	**(06 Hours)**
Interval Estimation. Confidence interval. Confidence level. Confidence interval for mean of normal population – implications for non-normal data. Confidence interval for variance of normal population. Confidence interval for difference of means of two independent normal population – when variances are equal and not equal. Confidence interval for difference of means of two paired normal population. Confidence interval for variance of normal population. Large and small sample confidence interval for proportions.		

References:

1. Medhi J. Statistical Methods: An Introductory Text. New Delhi: New Age International; 2006.
2. Mitzenmacher M, Upfal E. Probability and Computing. Cambridge University Press; 2005.
3. Mood AM, Graybill FA, Boes DC. Introduction to the Theory of Statistics. Auckland: McGraw-Hill; 1974.
4. Rohatgi VK, Saleh E. An Introduction to Probability and Statistics. Wiley Series in Probability and Statistics. Wiley; 2015.
5. Srivastava M, Khan A, Srivastava N. Statistical Inference Theory of Estimation. PHI Learning Pvt. Ltd; 2014.

Mapping of COs to POs

	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10
CO1	✓			✓						
CO2	✓			✓				✓		
CO3	✓			✓				✓		
CO4	✓			✓				✓		
CO5	✓									

CO6	✓			✓				✓		
CO7	✓			✓				✓		
CO8			✓	✓				✓		
CO9			✓	✓				✓		
CO10			✓	✓				✓		

Yenepoya (Deemed to be University) M.Sc. Data Science (2025) FS02DS-1C3: Principles of Data Science		
Teaching Scheme: Theory: 4 hrs/week		**Credits:** 4

Course Outcomes: Students will be able to

CO1: Define the core concepts and interdisciplinary nature of Data Science, including its scope and applications in various domains.

CO2: Describe the Data Science lifecycle, including data collection, analysis, interpretation, and deployment, and explain the importance of data transformation and cleaning techniques.

CO3: Identify different types of outliers and apply statistical methods for outlier detection in datasets.

CO4: Evaluate missing data scenarios in health research, discussing strategies for prevention and imputation techniques to handle missing data.

CO5: Explain the concepts of descriptive and inferential statistics, including measures of central tendency, dispersion, and shape, and their relevance in data analysis.

CO6: Conduct correlation analysis using Pearson and Spearman methods and apply Chi-square tests to evaluate relationships within datasets.

CO7: Apply various normalization techniques and scaling methods to prepare data for analysis, understanding the implications of each technique.

CO8: Utilize R and Python libraries for effective data transformation, including handling outliers and performing log and power transformations.

CO9: Apply hypothesis testing principles and interpret p-values, significance levels, and errors (Type I and Type II) to make informed decisions based on statistical results.

CO10: Select and perform appropriate parametric and non-parametric tests, such as t-tests, ANOVA, and Chi-square, to analyse data across different scenarios, interpreting outcomes to support research hypotheses.

Course Contents

Unit I	Introduction to Data Science	(10 Hours)
Definition, Scope, and Applications of Data Science,Interdisciplinary Nature of Data Science, Roles and Responsibilities of Data Scientist. Variables, Scales of measurement, Data: Definition, Types of data – Qualitative data and Quantitative data, Primary data and Secondary data,Data Science Lifecycle- Data collection, analysis, interpretation, and deployment. Data transformation and cleaning- Outlier Detection, Handling Missing Values, Data Scaling and Normalization.		
Unit II	**Data Management**	**(14 Hours)**
Outlier detection- definition of outlier, types of outliers. Statistical methods for dealing with outliers. Introduction to Missing Data: Definition of missing data, Importance of addressing missing data, Overview of types of missing data (MCAR, MAR, MNAR). Strategies for Handling Missing Data- Prevention Strategies: Data collection protocols to minimize missingness, Importance of study design. Imputation Techniques: Mean/mode/median imputation, Hotdeck imputation, Indicator method, Regression imputation, Multiple imputation, Maximum likelihood estimation. Sensitivity Analysis: Importance of testing assumptions.		
Unit III	**Data Description**	**(10 Hours)**
Basic Statistics: Definition, Types of statistics- Descriptive Statistics and Inferential Statistics, Measures of central tendency – Mean, Median, Mode. Measures of dispersion – Range, Variance, Standard Deviation, IQR, Coefficient of variation. Measures of shape – Skewness, Kurtosis. Correlation analysis Data Visualization –Line graphs, Bar chart, Pie chart, Histograms, Boxplots, Scatter plots.		
Unit IV	**Data Transformation**	**(10 Hours)**
Role of normalization in data preprocessing. Min-max scaling. Mean normalization. Z - score normalization. Decimal scaling. Logarithmic and power transformation. Square root transformation. BoxCox transformation. Normalization versus standardization. Handling outliers – winsorizing, trimming, robust scaling. Quantile transformation – transforming featured to follow a uniform or normal distribution. R and Python libraries for data transformation.		
Unit V	**Inference**	**(14 Hours)**

Hypothesis testing: Concept of statistical hypothesis and research hypothesis, errors in decisionmaking Type-I and Type II error, size and power of the statistical test, level of significance, pvalue and statistical significance. Statistical significance Vs Clinical significance.

Parametric Tests: Hypothesis testing for continuous data- one and two sample t test, paired t test, z- test for proportions, One-way ANOVA. Multiple Comparison tests, repeated measures ANOVA.

Non-Parametric Tests: Need and scope, sign test, median test, Mann Whitney U test, Wilcoxon Signed rank test, Kruskal Wallis test, Chi-square test of Independence of Attributes, two related sample comparison; its extension Willcoxon sign rank test, McNemar test and paired comparison. Friedman ANOVA and multiple comparison.

References:

1. Cielen D, Meysman A. Introducing Data Science. Simon and Schuster; 2016.
2. Gibbons JD, Chakraborti S. Nonparametric Statistical Inference. CRC Press; 2010.
3. Grus J. Data Science from Scratch: First Principles with Python. O'Reilly Media; 2019.
4. Ozdemir S. Principles of Data Science: Learn the Techniques and Math You Need to Start Making Sense of Your Data. Birmingham: Packt Publishing; 2017.
5. Pierson L. Data Science for Dummies. For Dummies; 2021
6. Srivastava M, Khan A, Srivastava N. Statistical Inference Testing of Hypotheses. PHI Learning Pvt. Ltd; 2014

Mapping of COs to POs

	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10
CO1	✓								✓	✓
CO2	✓								✓	✓
CO3	✓		✓	✓				✓		
CO4	✓		✓	✓				✓		

CO5	✓		✓	✓				✓		
CO6	✓		✓	✓				✓		
CO7	✓		✓	✓				✓		
CO8			✓	✓	✓	✓				✓
CO9			✓		✓	✓	✓	✓		✓
CO10			✓		✓	✓	✓	✓		✓

Yenepoya (Deemed to be University)

M.Sc. Data Science (2025)

FS02DS-1C4: Database Management Systems

Teaching Scheme:		**Credits:**
Theory: 4 hrs/week		4

Course Outcomes: Students will be able to

CO1. Describe the purpose and structure of Database Management Systems and data models.

CO2. Create Entity-Relationship (ER) models using entities, attributes, relationships, and keys.

CO3. Explain SQL concepts, including data types, DDL, DML, DCL, and TCL commands.

CO4. Perform basic database operations, including insertions, deletions, and updates using SQL.

CO5. Understand and implement advanced SQL features such as joins, views, and transactions.

CO6. Apply relational integrity constraints and create functions, procedures, and triggers in SQL.

CO7. analyse relational algebra and tuple relational calculus for query formulation.

CO8. Design and convert E-R and EER diagrams into relational tables, addressing design issues.

CO9. Discuss the principles of normalization and Codd's rules in relational database design.

CO10. Apply normalization techniques to achieve 2NF, 3NF, and BCNF using functional dependencies.

Course Contents

Unit I	**Introduction to Database System and ER Modelling**	**(12 Hours)**

Introduction to Database Management Systems, Purpose of Database Systems, View of Data, Database Languages, Database System Structure, Data Models, Database Design and ER Model: Entity, Attributes, Relationships, Constraints, Keys

Unit II	**Structured Query Language Basics**	**(10 Hours)**

SQL Overview: Data Types and Literals, DDL, DML, DCL, TCL. Data Definitions, Basic Structure Operations, Additional Operations, Set Operations, Null Values, Aggregate Functions, Nested Sub-Queries. Modifications of Database: Deletion, Insertion and Updates.

Unit III	**Intermediate Structured Query Language**	**(12Hours)**

Joins Expressions, Views, Transactions. Relational Integrity: Domain, Referential Integrities, Enterprise

Constraints. Data Types and Schemas, Authorizations, Functions and Procedures, Triggers		
Unit IV	**Relational Query Languages & Database Design**	**(10 Hours)**
Relational Algebra, Tuple relational Calculus, Design Process, Entity Relationship Model, ER Diagram, Design Issues, Extended E-R Features, converting E-R & EER diagram into tables.		
Unit V	**Relational Model Normalization**	**(12 Hours)**
Relational Database Design Relational Model: Basic concepts, Attributes and Domains, CODD's Rules, Database Design: Features of Good Relational Designs, Normalization, Atomic Domains and First Normal Form, Decomposition using Functional Dependencies, Algorithms for Decomposition, 2NF, 3NF, BCNF.		

References:

1. Chodorow K. MongoDB: The Definitive Guide. O'Reilly Media, Inc.; 2013.
2. Desai BC. An Introduction to Database Systems. Galgotia Publications; 1981.
3. Elmasri R, Navathe S. Fundamentals of Database Systems. Pearson Education; 2007.
4. Pakhira MK. Database Management System. PHI Learning Pvt. Ltd; 2012.
5. Silberschatz A, Korth HF, Sudarshan S. Database System Concepts. McGraw-Hill Science; 2006.

Mapping of Cos to POs

	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10
CO1		✓					✓	✓		
CO2		✓		✓	✓		✓	✓		
CO3		✓		✓	✓		✓	✓		
CO4				✓	✓		✓	✓		
CO5		✓			✓		✓	✓		
CO6		✓		✓	✓		✓	✓		

CO7		✓		✓	✓		✓	✓		
CO8		✓		✓	✓		✓	✓		
CO9		✓		✓	✓		✓	✓		
CO10		✓		✓	✓		✓	✓		

Yenepoya (Deemed to be University)

M.Sc. Data Science (2025)

FS02DS-2P1:

Practical I: Based on FS02DS-1C1 and FS02DS-1C3 using Python

Teaching Scheme:		**Credits:**
Practical: 6 hrs/week		3

Course Outcomes: Students will be able to

CO1. Familiarize with Python‘s syntax, data types, operators, and basic programming structure, enabling foundational coding skills.

CO2. Develop the ability to control program flow using loops and conditional statements, and work with mathematical and random functions for dynamic programming.

CO3. Gain proficiency in creating user-defined functions, modules, and packages, fostering modular and reusable code.

CO4. Understand and implement file operations to read, write, and manipulate files and directories, ensuring effective data management.

CO5. Apply OOP concepts such as encapsulation, inheritance, and polymorphism in Python to build structured and scalable programs.

CO6.Explore array-based data manipulation, vectorized computation, and mathematical operations, setting a strong foundation for numerical data handling.

CO7.Develop skills in advanced indexing, slicing, transposing, and Boolean indexing for efficient array manipulation.

CO8.Master universal functions for element-wise operations and statistical methods to handle complex datasets.

CO9.Work with pandas‘ Series and DataFrame structures to organize, filter, and map data efficiently.

CO10.Summarize and analyse data using descriptive statistics and handle data import/export operations for seamless integration

C011.Create and manage structured databases, tables, and constraints to establish a reliable data

storage foundation.

C012.Perform essential SQL operations such as INSERT, UPDATE, and DELETE, and apply complex calculations to data within databases.

C013.Implement SQL CASE statements, joins, and date/time functions for versatile data querying.

CO14.Use SQL to handle multiple tables, analyze time-based changes, and generate analytical reports.

CO15.Employ advanced SQL features, including SET operators, GROUP BY enhancements, and subqueries for comprehensive data handling.

CO16.Prepare and manipulate machine learning datasets using SQL to build familiarity with database-oriented data preparation.

Course Contents

List of Python Practical

1. Introduction to Python - Features, Identifiers, Reserved words, Indentation, Comments,Built- in Data types and their Methods: Strings, List, Tuples, Dictionary, Set – Type Conversion-Operators.
2. Decision Making- Looping- Loop, Control statement- Math and Random numberfunctions.
3. User defined functions – function arguments & its types, User defined Modules and, Packages in Python
4. Files: File manipulations, File and Directory related methods.
5. Python Exception Handling. OOPs Concepts – Class and Objects, Constructors – Data hiding-Data Abstraction- Inheritance, Types of Inheritance, Polymorphism, Encapsulation, Anonymous Function.
6. NumPy Basics: Arrays and Vectorized, Computation- The NumPy, ndarray- Creatingndarrays-Data Types for ndarrays- Arithmetic
7. Arrays- Basic Indexing and Slicing – Boolean IndexingTransposing Arrays with NumPy.
8. Universal Functions: Fast Element-Wise Array Functions- Mathematical and Statistical Methods-Sorting Unique and Other Set Logic.

9. Introduction to pandas Data Structures: Series, Data Frame, Essential Functionality: Dropping Entries Indexing, Selection, and Filtering- Function Application andMapping- Sorting and Ranking.
10. Summarizing and Computing Descriptive Statistics- Unique Values, Value Counts, and Membership. Reading and Writing Data in Text Format.

List of SQL Practical

1. Creating and Managing Structured Database and Tables: Create Database, Create Table, perform other DDL operations on table, Handle different constraints on the table
2. SQL Statements I:Using INSERT, UPDATE, DELETE operations on tables, B. Basic SQL Statements, Restricting and sorting data, Simple inline calculations, Complex, inline calculations
3. SQL Statements II:CASE Statement in SQL query, implementing different types of join, Performing various, Date/Time functions on SQL queries
4. Exploratory Data Analysis with SQL: Handling multiple tables simultaneouslyAnalysing changes over time in different tables Analytical reporting using SQL, Create and manage view, Create and manage CTE
5. Advanced Query Handling: A. Using SET Operators, Datetime Functions, Enhancements to the GROUP BY Clause, Advanced Subqueries
   Create and handle any two Machine Learning Datasets using simple SQL

Mapping of Cos to POs

	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10
CO1			✓	✓	✓	✓	✓		✓	✓
CO2			✓	✓	✓	✓	✓		✓	✓
CO3			✓	✓	✓	✓	✓		✓	✓
CO4			✓	✓	✓	✓	✓		✓	✓

CO5			✓	✓	✓	✓	✓		✓	✓
CO6			✓	✓	✓	✓	✓		✓	✓
CO7			✓	✓	✓	✓	✓		✓	✓
CO8			✓	✓	✓	✓	✓		✓	✓
CO9			✓	✓	✓	✓	✓		✓	✓
CO10			✓	✓	✓	✓	✓		✓	✓
CO11			✓	✓	✓	✓	✓		✓	✓
CO12			✓	✓	✓	✓	✓		✓	✓
CO13			✓	✓	✓	✓	✓		✓	✓
CO14			✓	✓	✓	✓	✓		✓	✓
CO15			✓	✓	✓	✓	✓		✓	✓
CO16			✓	✓	✓	✓	✓		✓	✓

Yenepoya (Deemed to be University)

M.Sc. Data Science (2025)

FS02DS-2P2:

Practical II: Data Analysis Using R

Teaching Scheme:		**Credits:**
Practical: 6 hrs/week		3

Course Outcomes: Students will be able to

CO1: Use R as a simple calculator

CO2. Perform vector and matrix operations in R

CO3. Choose and perform parametric tests and interpret the results.

CO4. Choose and perform non parametric tests and interpret the results.

CO5. Detect and manage the outliers using suitable statistical techniques

CO6. Detect and impute the missing values using suitable statistical techniques.

Course Contents

List of Practical I: Based on Statistical Foundation for Data Science using R:

1. Handling scalars, vectors, matrices, lists and data frames
2. Data Management and Visualization
3. Parametric Inference
4. Non-Parametric Inference
5. Detecting outliers
6. Handling missing values.

Mapping of Cos to POs

	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10
CO1			✓	✓	✓	✓	✓			✓
CO2			✓	✓	✓	✓	✓			✓
CO3			✓	✓	✓	✓	✓		✓	✓
CO4			✓	✓	✓	✓	✓		✓	✓

CO5			✓	✓	✓	✓	✓		✓	✓
CO6			✓	✓	✓	✓	✓		✓	✓

Semester II

Yenepoya (Deemed to be University) M.Sc. Data Science (2025) FS02DS-2E1: Essential Statistics for Research		
Teaching Scheme: Theory: 3 hrs/week		**Credits:** 3

Course Outcomes: Students will be able to

CO1. Define data and variables, identify different types of data, and provide examples

CO2. Calculate and interpret measures of central tendency and measures of dispersion.

CO3. Use Bayes' Theorem in examples related to diagnostic tests and Receiver Operating Characteristic (ROC) curves.

CO4. Define and utilize the Binomial and Normal distributions, provide examples of their applications in healthcare research, and establish reference ranges using probability distributions.

CO5.Identify different types of random sampling methods, apply them to healthcare scenarios.

CO6.Explain the concept of confidence intervals, their rationale, and their applications in healthcare research.

CO7. Calculate confidence intervals for population means and population proportions.

CO8. Calculate and interpret various measures of relationship and give examples.

Course Contents

Unit I	Descriptive Methods	(10 Hours)

Role of Statistics in evidence-based medicine/health care (EBM/EBHC): some examples from literature.

Data and Variables: definition, types, and examples.

Presentation of Data: Tabular and Graphical Methods,

Measures of Central Tendency:mean, median, mode, geometric mean, merits, and demerits

Measures of Dispersion:standard deviation, inter quartile range, coefficient of variation

Measures of Shape:skewness and kurtosis

Exploratory Data Analysis:selection of appropriate summary measure, outlier detection, deviation

from normality, subgroup analysis, homogeneity of variables among groups.

Unit II	Application of theory of Probability	(11 Hours)

Probability and associated measures: random experiment, sample space, events, types of events, probability of Events subjective, classical, and frequentist approach in probability, applications, and limitations, conditional probability, risk, rates, odds, risk ratio and odds ratio.

Bayes Theorem and Applications: Statement, examples, accuracy measures (Sensitivity, Specificity, PPV NPV) of a diagnostic test, diagnostic cut off using a Receiver Operating Characteristic (ROC) curve.

Binomial and Normal Distribution: Definition, examples, application, reference ranges using a probability distribution

Unit III	Sampling and Estimation	(11 Hours)

Terminology: population and sample, parameter and statistic, example from research, random sample and non-random sample. Types of Random Sampling: simple random sampling, stratified sampling, systematic sampling, cluster sampling.

Confidence interval: definition, rationale, interpretation, examples, application.

Confidence interval for population mean: definition, standard error of mean, examples, application and limitation Confidence interval for population proportion: definition, standard error of proportion, examples, application and limitation.

Sample Size Estimation: Estimation of mean and prevalence

Unit IV	Summarizing the Relationship	(10 Hours)

Correlation Coefficient: linear and monotonic correlation, Pearson correlation coefficient, Spearman rank correlation, definition, examples, limitation, scatter plot.

Summarizing Nominal Association: contingency coefficient, Phi coefficient, Cramer‘s V

Summarizing Ordinal Association: definition, rationale, examples, application.

Kappa Measure of Agreement: definition, confidence interval for kappa, application, limitations

References:

1. Medhi J. Statistical Methods: An Introductory Text. New Delhi: New Age International; 2006.

2. Mood AM, Graybill FA, Boes DC. Introduction to the Theory of Statistics. Auckland: McGraw-Hill; 1974.
3. Rohatgi VK, Saleh E. An Introduction to Probability and Statistics. Wiley Series in Probability and Statistics. Wiley; 2015.
4. Srivastava M, Khan A, Srivastava N. Statistical Inference Theory of Estimation. PHI Learning Pvt. Ltd; 2014.

Yenepoya (Deemed to be University)

M.Sc. Data Science (2025)

FS02DS-2C1: Multivariate Analysis

Teaching Scheme: Theory: 4 hrs/week		**Credits:** 4

Course Outcomes: Students will be able to

CO1. Organize and describe multivariate data using mean vectors, covariance matrices, and graphical tools

CO2. Apply the multivariate normal distribution, estimate parameters, and assess normality in real data.

CO3. Understand and apply statistical tests like Hotelling's T^2 and MANOVA for making inferences about population means.

CO4. Discuss the basic concepts of multivariate linear regression and interpret results from various multivariate tests.

CO5.Explain the procedures for discriminant analysis, including the use of linear and quadratic discriminant functions, and discuss assumptions about variance-covariance matrices.

CO6. Interpret classification outcomes and discuss methods for evaluating misclassification rates.

CO7. Describe various clustering methods, and explain criteria for choosing the optimal number of clusters.

CO8.Explain canonical correlation, including the calculation of canonical variates and interpretation of results, with applications in multivariate contexts.

CO9. Describe the objectives and theoretical basis of principal component analysis (PCA) and explain the interpretation of principal components.

CO10. Explain factor analysis methods, including factor loadings and rotations, and discuss its relationship to PCA.

Course Contents

Unit I	**Introduction to Multivariate Analysis**	**(14 Hours)**

Unit 1: Introduction: Organization of multivariate data and descriptive statistics(meanvectors, variance covariance matrix and their properties), Graphical andpictorialrepresentation of multivariate data-scatter plot matrix, stars, Chernoff faces.Multivariate Normal distribution: density function and its properties, maximumlikelihoodestimation, detection of outliers, connection to Mahalanobis distance,geometry of multivariate normal distribution, assessing the assumption of multivariatenormality, transformations to near normality, Demonstration using medical data.		
Unit II	**Inference regarding multivariate population mean:**	**(10 Hours)**
Hotelling's T squaredistribution- single, paired, two independent sample inference, One Way MultivariateAnalysis of Variance (MANOVA), Wilk's lambda, Roy's test, Pillai and Lawley trace,Simultaneous confidence interval for treatment effects, Bonferroni method of multiplecomparisons,Introduction to basics of multivariate linear regression models.		
Unit III	**Discriminant Analysis:**	**(10 Hours)**
Procedure for discriminant analysis, linear andquadratic discriminant functions for two and several populations, Bartlett's test forhomogeneity of variance covariance matrices, quadratic discriminant analysis,estimating misclassification probabilities- confusion table, robustness to multivariatenormality assumption, classification trees.		
Unit IV	**Cluster Analysis:**	**(10 Hours)**
similarity measures, hierarchical clustering methods- singlelinkage, complete linkage, average linkage, Ward's method; non-hierarchicalclustering methods- k means method, choosing the optimal number of clusters.Canonical correlation: canonical variates and canonical correlation, population andsample canonical correlations, tests of significance, application in health sciences.		
Unit V	**Principal Component Analysis:**	**(12 Hours)**
Goal of principal component analysis,population principal components, sample principal components, number of principalcomponents to retain, interpretation, standardization of the data, graphicalpresentation of components, large sample inference.Factor Analysis: Orthogonal factor model, model assumptions, the principalcomponent method, maximum likelihood estimation, rotation-orthogonal rotation, obliquerotations, interpretation, , communalities, estimation of factor loading		

factor scores, validity ofthe factor analysis model, the relationship of factor analysis to principal component analysis.

References:

1. Anderson TW. An Introduction to Multivariate Statistical Analysis. Wiley; 2003.
2. Dillon WR, Goldstein M. Multivariate Analysis. John Wiley & Sons; 1984.
3. Johnson RA, Wichern DW. Applied Multivariate Statistical Analysis. Pearson Prentice Hall; 2007.
4. Maxwell AE. Multivariate Analysis in Behavioural Research. Chapman and Hall; 1977.
5. Raykov T, Marcoulides GA. An Introduction to Applied Multivariate Analysis. Routledge; 2008.
6. Rencher AC, Christensen WF. Methods of Multivariate Analysis. Wiley; 2012.
7. Tacq J. Multivariate Analysis Techniques in Social Science Research: From Problem to Analysis. London: Sage; 2007.

Mapping of COs to POs

	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10
CO1	✓			✓				✓		
CO2	✓		✓	✓				✓		
CO3	✓		✓	✓				✓		
CO4	✓							✓		
CO5	✓							✓		
CO6	✓		✓	✓				✓		
CO7	✓		✓	✓				✓		
CO8	✓		✓	✓				✓		
CO9	✓		✓	✓				✓		

CO10	✓		✓	✓				✓		

Yenepoya (Deemed to be University)

M.Sc. Data Science (2025)

FS02DS-2C2: Machine Learning

Teaching Scheme:		**Credits:**
Theory: 4 hrs/week		4

Course Outcomes: Students will be able to

CO1. Explain the fundamental concepts of machine learning, including its types (supervised, unsupervised, and reinforcement learning) and illustrate with real-world applications.

CO2. Identify different types of machine learning through case studies.

CO3. Apply supervised learning techniques to solve classification problems and evaluate model performance.

CO4. Implement regression trees, decision trees, and random forests, and compare the effectiveness of ensemble techniques like bagging and boosting in improving model accuracy.

CO5. Distinguish between various unsupervised learning methods and implement clustering algorithms.

CO6. Analyse the quality of clusters using evaluation metrics, and apply association rules and genetic algorithms to discover patterns in data.

CO7.Describe the architecture of neural networks and explain the functionality of perceptron and multilayer perceptron in solving non-linear problems.

CO8.Implement the backpropagation algorithm with various optimization techniques to train networks effectively and prevent overfitting by tuning the network structure.

C09.Demonstrate model combination techniques like voting, bagging, boosting, and stacking to enhance predictive performance, and compare their effectiveness in diverse datasets.

CO10. Examine the trade-off between diversity and accuracy in ensemble learning and apply reinforcement learning principles to adapt models to dynamic environments.

Course Contents

Unit I	**Introduction to Machine Learning**	**(06 Hours)**

What is machine learning? Examples of machine learning applications. Types of machine learning - supervised and unsupervised – definition and examples. Reinforcement learning. Case studies.		
Unit II	**Supervised Learning**	**(16 Hours)**
Introduction to machine learning, Supervised Learning: Introduction tosupervisedlearning, K – nearest neighbourhood algorithm, Naïve Bayes and Bayesian networks.SupportVectorMachines and Kernel Methods, Optimal Separating Hyperplane, Soft-MarginClassifier. Regression Tree, CART, Ensemble techniques: Decision Tree and RandomForest, Model evaluation techniques.		
Unit III	**Unsupervised Learning**	**(14 Hours)**
Unsupervised Learning: Introduction to unsupervised learning; Partitioning methods:K-Means, K-Medoid, Hierarchical methods: BIRCH; Density-based methods:DBSCAN, OPTICS, DENCLUE; Grid-based methods: STING, CLIQUE; Measuring clustergoodness. Association rules, Genetic algorithms.		
Unit IV	**Multilayer Perceptron**	**(10Hours)**
Understanding the brain. Neural networks. The perceptron. Learning Boolean functions. The multilayer perceptron. Backpropagation algorithm. Multiple hidden layers. Training procedures – improving convergence, momentum, adaptive learning rate. Overtraining. Structuring the network. Tuning the network size. Bayesian view of learning.		
Unit V	**Combining Multiple Learners and Reinforcement Learning**	**(10 Hours)**
Introduction. Diverse learners – different algorithms, different hyperparameters, different input representations, different training sets. Diversity vs accuracy. Model combination schemes - Mult expert combination, multistage combination. Voting. Bagging. Boosting. Stacking. Fine tuning an ensemble.		

References:

1. FlachP, Machine Learning: The Art and Science of Algorithms that make sense of data, Cambridge University Press, 1st Edition, 2012
2. Alpaydin E, Introduction to Machine Learning, PHI, 2nd edition, 2013

3. Murphy K, Machine Learning: A Probabilistic Approach, MIT Press, 1st Edition, 2012
4. BishopCM, Pattern Recognition and Machine learning, Springer, 1st Edition, 2013

Mapping of COs to POs

	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10
CO1		✓		✓	✓		✓	✓		
CO2		✓		✓	✓		✓	✓		
CO3		✓		✓	✓		✓	✓		
CO4		✓		✓	✓		✓	✓		
CO5		✓		✓	✓		✓	✓		
CO6		✓		✓	✓		✓	✓		
CO7		✓		✓	✓		✓	✓		
CO8		✓		✓	✓		✓	✓		
CO9		✓		✓	✓		✓	✓		
CO10		✓		✓	✓		✓	✓		

Yenepoya (Deemed to be University)

M.Sc. Data Science (2025)

FS02DS-2C3: Applied Regression Analysis

Teaching Scheme:		**Credits:**
Theory: 4 hrs/week		4

Course Outcomes: Students will be able to

CO1. Explain the foundational assumptions and methodology for least square estimation in simple linear regression, including standard error calculations.

CO2. Interpret the significance of model parameters (slope, intercept) through hypothesis testing and evaluate model fit using the coefficient of determination and ANOVA.

CO3. Describe the structure and assumptions of multiple linear regression, estimating parameters and evaluating model fit using R^2 and Adjusted R^2

CO4. Explain the role of dummy variables in regression models and conduct residual analysis to assess model adequacy.

CO5. Explain the Gauss-Markov assumptions and evaluate heteroscedasticity, multicollinearity, and autocorrelation in general linear models.

CO6. Apply statistical tests, such as the Durbin-Watson test, to detect autocorrelation and describe methods to manage heteroscedasticity and multicollinearity in regression models.

CO7. Summarize methods for variable selection, including forward selection, backward elimination, and stepwise regression.

CO8. Explain the use of information criteria, particularly the Akaike Information Criterion (AIC), for model selection and assess residuals to diagnose model performance.

CO9.Differentiate between linear and nonlinear regression models, discussing their transformations and limitations in applied settings.

CO10.Outline the principles of logistic regression for categorical outcomes, interpret model coefficients, and explain adjustments for confounding and interaction factors.

Course Contents

Unit I	Simple Linear Regression	(8 Hours)
Introduction to regression analysis, Assumptions and least square estimation of model parameters, Standard error of the estimators, testing of hypothesis on the significance of slope and intercept. Coefficient of determination, residual analysis, ANOVA.		
Unit II	**Multiple Linear Regression Models**	**(10 Hours)**
Multiple linear regression: Description of the model. Estimation of regression parameters and error variance, properties, test of hypothesis of regression parameters, Significance of regression (ANOVA, R2, Adjusted R2), dummy variable regression-general concepts and uses. Residual analysis.		
Unit III	**General Linear Models**	**(14 Hours)**
General linear models: Introduction- Gauss Markov set up, assumptions- homoscedasticity and heteroscedasticity, testing for heteroscedasticity, generalised least squares estimator, multicollinearity, its detection and methods to deal with multicollinearity. Autocorrelation, Durbin- Watson test, estimation of model parameters.Variance stabilising transformations to linearize the model. Analytical methods for selecting a transform		
Unit IV	**Feature EngineeringandResidual Analysis**	**(10 Hours)**
Selection of variables, forward selection, backward elimination and stepwise regression (Algorithms only), Information Criteria, Akaikeinformation criteria, BIC, Residual analysis		
Unit V	**Advanced Regression Models**	**(14 Hours)**
Introduction to non-linear regression. Nonlinear regression- Transformation to a linear model, Usefulness of the nonlinear regression method, limitations of the nonlinear regression method, use of re-sampling procedures in regression. Logistic Regression: Introduction, logistic regression models- Dichotomous, Polychromous and Continuous independent variables. Fitting logistic regression model, Conditional Logistic Regression, Interpreting coefficients. Testing the significance of model. Adjustment for Confounding and interaction factors.		

References:

1. Draper N.R. and Smith H, Applied Regression Analysis. 3rdEd. John Wiley andSons. Inc;1998
2. Hosmer D W and Lemeshow S, Applied Logistic Regression. Wiley;1989
3. Johnston J, Econometric Methods. 3rd Ed. McGraw Hill;1984
4. Johnston J. and Dinardo J, Econometric Methods. 4thEd. McGraw-HillCompanies;1997
5. Montgomery, D. C., Peck, Introduction to linear regression analysis. John Wiley andSons. 3rdEd; 2003
6. Weisberg S, Applied Linear Regression Wiley;2005
7. Zar J H, Bio statistical Analysis. Pearson Education. 4 edition;2006

Mapping of COs to POs

	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10
CO1	✓							✓		
CO2	✓		✓	✓				✓		✓
CO3	✓							✓		
CO4	✓		✓	✓				✓		✓
CO5	✓							✓		
CO6	✓		✓	✓				✓		✓
CO7	✓							✓		
CO8	✓		✓	✓				✓		✓
CO9	✓							✓		
CO10	✓		✓	✓				✓		✓

Yenepoya (Deemed to be University)

M.Sc. Data Science (2025)

FS02DS-2C4: Data warehousing and Data Mining

Teaching Scheme:		**Credits:**
Theory: 4 hrs/week		4

Course Outcomes: Students will be able to

CO1. Explain the need for data warehousing in organizations and how it addresses strategic information requirements.

CO2. Analyse the evolution of business intelligence and describe the building blocks of data warehousing, including architectures and metadata.

CO3. Describe the principles of dimensional modelling and the advantages of different schema types, including star and snowflake schemas.

CO4. Demonstrate the ETL process, including data extraction, transformation, and loading, using practical examples.

CO5. Apply dimensional modelling techniques to design data warehouses and data marts, demonstrating the use of aggregate fact tables and miscellaneous dimensions.

CO6.Evaluate and implement various ETL integration approaches, discussing their effectiveness in data warehousing.

CO7. Identify the different users of information in a data warehouse and their specific information needs.

CO8.Explain the demand for OLAP in data warehousing, including major features, functions, and implementation considerations.

CO9.Define data mining and its role as a process of knowledge discovery, highlighting major methodologies and trends in the field.

CO10. Apply data mining techniques, such as classification and clustering, to solve real-world problems and analyse complex data types.

Course Contents

Unit I	**Overview and Concept**	**(10 Hours)**

Need for Data warehousing: Escalating need for strategic information, failure of past system, operational versus decision support systems, data warehouse defined, data warehousemovement, Evolution of Business Intelligence. Data ware housing building blocks: Data ware house and Data Marts, architecture, components, metadata Trends in warehousing: Continue growth in data warehousing, significant trends		
Unit II	**Architecture and Infrastructure**	**(12 Hours)**
Principles of dimension modelling: Dimensional modelling basics, star schema, advantages, examples. Advanced dimension modelling: updates to dimensional tables, miscellaneous dimensions, snowflake schema, Aggregate Fact tables, families of Stars. Data Extraction, Transformation, and Loading: ETL overview, requirement, Data extraction, DataTransformation, Data Loading, other integration approaches		
Unit III	**Design and Data Preparation**	**(12 Hours)**
Advanced dimension modeming: updates to dimensional tables, miscellaneous dimensions, snowflake schema, Aggregate Fact tables, families of Stars. Data Extraction, Transformation, and Loading: ETL overview, requirement, Data extraction, DataTransformation, Data Loading, other integration approaches		

Unit IV	**Information Access and delivery**	**(10 Hours)**
Users of Information: Information from the data ware house, who will use information, information delivery, delivery tools OnLine Analytical Processing in the data warehouse: Demand for OLAP, major features and functions, OLAP models, OLAP implementation considerations Data ware house and the web: Web enables data warehouse, web-basedinformation delivery, OLAP and the WEB, building a web-enabled data warehouse		
Unit V	**Data Mining**	**(12 Hours)**
Why data mining, what is Data Mining, Data mining as a process of Knowledge Discovery,Major issues in Data Mining, Mining Frequent Pattern, Data Mining Trends and Research Frontiers: Mining Complex Data Types, other Methodologies of Data Mining, Data MiningApplications, Data Mining trends. Physical design process: Physical design steps, considerations, physical storage, indexing, performance enhancement techniques. Data warehouse development: Data warehouse testing,major deployment activity, security, backup and recovery. Growth and Maintenance: Monitoring the data warehouse, user		

training and support, managing the data warehouse.

References:

1. A M. Mining the Social Web: Analyzing Data from Facebook, Twitter, LinkedIn, and Other Social Media Sites. O'Reilly; 2011.
2. Anahory S, Murray D. Data Warehousing in the Real World: A Practical Guide for Building Decision Support Systems. Addison-Wesley; 1997.
3. Guptha G. Introduction to Data Mining with Case Studies. PHI Learning Private Limited; 2015.
4. Han J, Kamber M. Data Mining: Concepts and Techniques. Elsevier/Morgan Kaufmann; 2012.
5. Ponniah P. Data Warehousing Fundamentals: A Comprehensive Guide for IT Professionals. Wiley; 2006.

Mapping of COs to POs

	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10
CO1		✓						✓		✓
CO2		✓						✓		✓
CO3		✓						✓		✓
CO4		✓	✓	✓	✓		✓	✓		✓
CO5		✓	✓	✓	✓		✓	✓		✓
CO6		✓	✓	✓	✓		✓	✓		✓
CO7		✓	✓	✓	✓		✓	✓		✓
CO8		✓	✓	✓	✓		✓	✓		✓
CO9		✓						✓		✓
CO10		✓	✓	✓	✓		✓	✓		✓

Yenepoya (Deemed to be University)

M.Sc. Data Science (2025)

FS02DS-2P1:

Practical III: Based on FS02DS-1C1 and FS02DS-1C3 using Python

Teaching Scheme:		**Credits:**
Practical: 6 hrs/week		3

Course Outcomes: Students will be able to

CO1. Assess and validate the assumption of multivariate normality in datasets using Q-Q plots and Chi-Square plots, ensuring the appropriateness of subsequent statistical analyses.

CO2. Demonstrate proficiency in applying Hotelling's T^2 statistic for both one-sample and two-sample problems, interpreting the results to draw conclusions about multivariate means in various contexts.

CO3. Conduct and interpret Multivariate Analysis of Variance (MANOVA), utilizing overall significance tests (Wilks' Lambda, Roy's test, Pillai's trace, and Lawley‘s test).

CO4. Perform Principal Component Analysis (PCA), Factor Analysis, and Canonical Correlation, interpreting the outcomes.

CO5. Construct and interpret simple and multiple linear regression models, applying appropriate methods for model selection and conducting thorough residual analyses to validate model assumptions.

CO6. Engage in feature engineering practices, including variable selection methods and regularization techniques (e.g., Ridge and Lasso), to improve model performance and interpretability while addressing issues of multicollinearity and overfitting.

CO7. Analyse mixed-effect models, understanding the significance of random effects in regression analyses.

Course Contents

List of multivariate practical:

1. Assessing multivariate normality –Q-Q plot and Chi-Square plot.
2. Hotellings T 2 statistic – I (one sample and two sample problem)
3. Hotellings T 2 statistic – II (Two independent samples problem)
4. MANOVA
5. Overall significance test of MANOVA- Wilk‘s lamda, Roy‘s test, Pillai and Lawley
6. Principal Component Analysis
7. Factor Analysis
8. Canonical Correlation
9. Cluster Analysis
10. Classification and discrimination

List of Regression Analysis practical:

1. Simple Linear Regression
2. Multiple Linear Regression
3. Detecting problems in the Regression Model – multicolinearity, heteroscedasticity, autocorrelation.
4. Residual Analysis
5. Feature engineering (variable selection methods)
6. Regularization techniques.
7. Generalized Linear Regression

Mapping of COs to POs

	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10
CO1			✓	✓	✓	✓	✓	✓	✓	✓
CO2			✓	✓	✓	✓	✓	✓	✓	✓
CO3			✓	✓	✓	✓	✓	✓	✓	✓
CO4			✓	✓	✓	✓	✓	✓	✓	✓

CO5			✓	✓	✓	✓	✓	✓	✓	✓
CO6			✓	✓	✓	✓	✓	✓	✓	✓
CO7			✓	✓	✓	✓	✓	✓	✓	✓

Yenepoya (Deemed to be University)

M.Sc. Data Science (2025)

FS02DS-2P2:

Practical IV: Based on FS02DS-2C2 and FS02DS-2C4 using Python

Teaching Scheme:		**Credits:**
Practical: 6 hrs/week		3

Course Outcomes:

CO1. Implement machine learning algorithms (KNN, Naïve Bayes, SVM) to classify and cluster datasets effectively.

CO2. Perform clustering and association rule mining to identify patterns and relationships in data, enhancing analytical insights.

CO3. Design Star and Snowflake schemas for data warehousing and conduct exploratory OLAP operations to analyse data structures.

CO4. Execute ETL processes for data extraction, transformation, and loading, ensuring accurate and clean data integration in data warehouses.

CO5. Build and train a Multilayer Perceptron (MLP) and apply ensemble techniques to improve model performance in predictive analytics.

Course Contents

List of Machine Learning Practical:

1. Introduction to Machine Learning and Data Preprocessing
2. Implementing K-Nearest Neighbors (KNN) for Classification
3. Building a Naïve Bayes Classifier and Bayesian Network
4. Support Vector Machine (SVM) and Kernel Methods
5. Building Decision Trees and Random Forest Classifiers
6. Clustering with K-Means, DBSCAN, and Hierarchical Clustering
7. Implementing Association Rule Mining
8. Building and Training a Multilayer Perceptron (MLP)
9. Ensemble Learning Techniques: Bagging, Boosting, and Stacking

List of MySQL practical.

1. Designing a Star Schema for a Data Warehouse

2. Implementing a Snowflake Schema
3. Exploratory OLAP Operations
4. Data Mining with Association Rules (Market Basket Analysis)

List of ETL practical

1. Data Extraction and Loading from Multiple Sources
2. Data Transformation with Cleansing and Filtering
3. Data Aggregation and Summarization
4. Loading Transformed Data into a Data Warehouse
5. Incremental Data Load and Change Data Capture

Mapping of COs to POs

	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10
CO1			✓	✓	✓	✓	✓	✓	✓	✓
CO2			✓	✓	✓	✓	✓	✓	✓	✓
CO3			✓	✓	✓	✓	✓	✓	✓	✓
CO4			✓	✓	✓	✓	✓	✓	✓	✓
CO5			✓	✓	✓	✓	✓	✓	✓	✓