Mahatma Education Society's

Pillai HOC College of Arts, Science & Commerce (Autonomous)
Rasayani

Affiliated to University of Mumbai
NAAC Accredited with "A+" Grade in cycle II
ISO 9001:2015 Certified

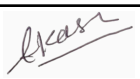
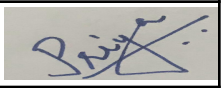
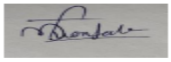
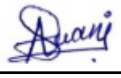
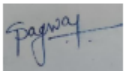
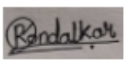
SYLLABUS

Bachelors of Science (B. Sc.) in Data Science
S. Y. B. Sc. Data Science

As per National Education Policy 2020

Academic Year 2026-27

Mahatma Education Society's
College Code: 870
PILLAI HOC COLLEGE OF ARTS, SCIENCE & COMMERCE
Pillai HOCL Educational Campus, HOC Colony, Rasayani, Via. Panvel, Dist. Raigad. Pin 410207
Tel: 02192 - 669000 / 01 / 02 / 03 / 04 / 05 / 06 / 07 / 08 / 09 Website : www.phcasc.ac.in Email : phcasc@mes.ac.in
(NAAC Accredited 'A+' Grade , CGPA – 3.26 in Cycle 2 & ISO 9001:2015 Certified)
Affiliated to the University of Mumbai, Approved by Government of Maharashtra
(AUTONOMOUS COLLEGE)

Sr. No	Name	Designation	Signature
1	Dr. Swapna Kadam	Vice Chancellor Nominee	
2	Dr. Annie Rajan,	Subject Expert	
3	Dr. Homraj Patelpaik	Subject Expert	
4	Mr. Swapnil H. Patil	Industry Representative	
5	Mr. Akash Ghadge	Alumni Representative	
6	Dr. Rinkoo Shantnu	Principal	
7	Ms. Priyanka Sorte	Member	
8	Ms. Priya Prakash	Member	
9	Ms. Harshita Singh	Member	
10	Ms. Mrunal Wanjale	Member	
11	Ms. Arpita Kante	Member	
12	Ms. Anita Mhatre	Member	
13	Ms. Sangeeta Menon	Member	
14	Ms. Aarti Wani	Member	
15	Ms. Rutuja Madane	Member	
16	Ms. Sonali Dagwar	Member	
17	Ms. Rutuja Kondalkar	Member	

INTRODUCTION

A **Data Science** degree program is a dynamic educational pathway that equips students with a multidisciplinary skill set essential for navigating the intricacies of the data-driven world. Foundational courses in mathematics and statistics lay the groundwork, while programming skills in languages like Python and Rare are honed for data manipulation and analysis. The curriculum delves into machine learning techniques, covering both supervised and unsupervised learning, and explores big data technologies such as Hadoop and Spark. Students gain practical experience in applying these skills to real-world problems through capstone projects, ensuring they are well-prepared to address the challenges of data science in diverse industries. Furthermore, Data Science degree programs emphasize the ethical considerations surrounding data use and privacy. Students engage in discussions about responsible conduct in data science, addressing the societal implications of their work. The program typically culminates in the development of strong communication skills, with a focus on data visualization and effective presentation of findings to non-technical stakeholders. Through a combination of theoretical knowledge, practical experience, and ethical considerations, graduates of Data Science degree programs are well- positioned to make meaningful contributions in a data-driven world.

Programme Outcomes (POs)

PO. No.	PO Title	POs in brief
PO1	Fundamental Knowledge Acquisition	Graduates will demonstrate a comprehensive and foundational knowledge of their chosen discipline along with an awareness of interdisciplinary connections.
PO2	Critical Thinking and Analytical Reasoning	Graduates will be able to analyze complex problems, synthesize data from multiple sources (qualitative and quantitative), and employ logical reasoning to formulate well-supported conclusions and arguments.
PO3	Effective Communication	Graduates will exhibit proficiency in both written and oral communication, articulating ideas clearly, persuasively, and ethically to diverse audiences
PO4	Problem Solving	Graduates will possess the ability to identify, formulate, and design solutions for real-world problems in their professional or social contexts, applying relevant theoretical knowledge and practical skills.
PO5	Information and Digital Literacy	Graduates will demonstrate the capability to locate, evaluate, and effectively use information from various sources, and utilize modern tools and Information and Communication Technology (ICT) for professional and academic tasks.
PO6	Research Skills and Scientific Temperament	Graduates will develop a sense of inquiry and research methodology, including the ability to design experiments (where applicable), collect and analyze data, and interpret results while maintaining scientific rigor and intellectual honesty.
PO7	Ethical Reasoning and Professional Integrity	Graduates will recognize ethical dilemmas, commit to professional and academic ethics, and demonstrate an understanding of moral and social responsibilities in their personal and professional conduct.
PO8	Employability and Professional Skills	Graduates will acquire the necessary job-ready skills, managerial competencies, and professional values to secure gainful employment or pursue advanced education in their respective fields.
PO9	Environmental and Sustainability Consciousness	Graduates will understand the importance of environmental conservation and sustainable development, displaying responsibility toward ecological challenges and advocating for healthy environmental practices.
PO10	Life-Long Learning	Graduates will develop the capacity for independent and self-directed learning to continuously upgrade their knowledge and skills, enabling them to adapt to rapid technological and societal changes.
PO11	Civic and Social Responsibility	Graduates will act as responsible citizens with an informed awareness of constitutional values, engaging proactively in community development and addressing social needs.
PO12	Empathy and Social Intelligence	Graduates will be able to cultivate and demonstrate affective, interpersonal, social and emotional intelligence.

Programme Specific Outcomes (PSOs)

PSO. No.	PSO Title	PSOs in brief
PSO1	Analytical Modeling and Solution Design	Students can formulate and implement robust predictive models and analytical solutions to address complex domain-specific problems by selecting and optimizing appropriate statistical and machine learning algorithms
PSO2	Scalable Data System Competency	Students can architect and deploy scalable data pipelines and storage mechanisms capable of ingesting, cleaning, and processing massive, unstructured datasets efficiently using distributed computing frameworks and cloud technologies.
PSO3	Responsible Practice and Research	Students can critically evaluate the ethical implications of data science interventions—including algorithmic bias, data privacy, and societal impact—while demonstrating the ability to independently research and integrate emerging methodologies.
PSO4	Professional Communication and Adaptability	Students can translate complex technical findings into actionable business insights for non-technical stakeholders through effective data storytelling and visualization, while demonstrating adaptability and leadership within cross-functional teams.

Evaluation Pattern

Marking Code	Marking Scheme
A	50 Marks Semester End Exam, 50 Marks Continuous Assessment (distributed within 15 Marks Class Test, 15 Marks Presentation & Assignment, 10 Marks Online Quiz, 10 Marks Attendance & Class Participation)
B	50 Marks Semester End Exam
C	100 marks Continuous Assessment (distributed within 30 Marks Class Test, 30 Marks Presentation & Assignment, 30 Marks Online Quiz, 10 Attendance & Class Participation)
D	50 Marks of Continuous Assessment (distributed within 15 Marks Class Test, 15 Marks Presentation & Assignment, 10 Marks Online Quiz, 10 Marks Attendance & Class Participation)
E	50 Marks Practical Examination (distributed within 30 Marks Practical Module 1 & 2*, 10* Marks Journal, 10 Marks Viva)

Course Structure

Semester - III							
Course Code	**Course Type**	**Course Title**	**Theory/ Practical**	**Marks**	**Credits**	**Lectures /Week**	**Evaluation Pattern**
HUSDS208	Major	DATA STRUCTURE AND ALGORITHM DESIGN	Theory	100	2	2	A
HUSDS208P	Major-Practical	PRACTICAL(HUSDS208)	Practical	50	1	2	E
HUSDS209	Major	DATA MINING	Theory	100	2	2	A
HUSDS209P	Major-Practical	PRACTICAL(HUSDS209)	Practical	50	1	2	E
HUSDS210	Major	DATA WAREHOUSING	Theory	100	2	2	A
HUSDS210P	Major-Practical	PRACTICAL(HUSDS210)	Practical	50	1	2	E
HUSDS211	Minor	PROBABILITY THEORY WITH DISCRETE DISTRIBUTIONS	Theory	100	2	2	A
HUSDS211P	Minor-Practical	PRACTICAL(HUSDS211)	Practical	50	1	2	E
HUSDS212	SEC	SCALA FOR DATA SCIENCE	Theory	100	2	2	A
HUSDS212P	SEC Practical	PRACTICAL(HUSDS212)	Practical	50	1	2	E
	AEC	हिंदी भाषा एवं साहित्य संवर्धन	Theory	50	2	2	D
	OE	RESEARCH METHODOLOGY	Theory	100	3	3	C
	CC	EXTENSION/NSS	Theory	50	2		D
				950	**22**		

Abbreviations:

SEC: Skill Enhancement Course
AEC: Ability Enhancement Course
VAC: Value Added Course
VEC: Value Education Course
IKS: Indian Knowledge System
OE: Open Elective

SEMESTER III

BOS	**Mathematics, Statistics and Computer Application**				
Course	**Data Structure and Algorithm Design**				
Course Code	**HUSDS208**	**Level**	**5.0**		
		Type	**Theory**	**Practical**	**Total**
Semester	**III**	**Credits**	2	1	3
Type	**Major**	**No of Teaching Hours**	30	30	60
Evaluation Pattern	**Total Marks**	**Semester End**	**Continuous**		**Practical**
	150	50	50		50

Learning Objectives	
1	To introduce the fundamental concepts of algorithms, their properties, and the importance of efficient algorithm design.
2	To develop a deep understanding of data structures such as linked lists, stacks, queues, trees, and graphs to support algorithmic problem-solving.
3	To teach mathematical techniques for analysing algorithm efficiency using asymptotic notations and mathematical analysis of recursive and non-recursive algorithms.
4	To equip students with the ability to apply efficient algorithms for sorting, searching, and graph traversal in real-world scenarios.
5	To enable students to compare, select, and justify appropriate algorithms and data structures based on time and space complexity requirements

Course Outcomes	
CO1	Understand and explain the fundamental principles and properties of algorithms and data structures, and apply them to solve computational problems.
CO2	Analyse the time and space complexity of algorithms using asymptotic notations.
CO3	Implement and apply various data structures such as linked lists, stacks, queues, trees, and graphs for solving different computational problems
CO4	Design and implement efficient algorithms for searching, sorting, and graph-based operations, evaluating their comparative performance in terms of complexity.

Modules at Glance

Module No.	**Content**	**No. of Hours**	**CO Mapping**
1	Introduction, Fundamentals of the Analysis of Algorithm Efficiency, Linked Lists, Stacks, Queues	15	CO1, CO2, CO3
2	Trees, Graph Algorithm, Sorting, Searching	15	CO3, CO4

Syllabus

Module No.	Content	No. of Lectures
1	**Introduction**: What Is an Algorithm, Fundamentals of Algorithmic Problem Solving, Important Problem Types, Fundamental Data Structures. **Fundamentals of the Analysis of Algorithm Efficiency:** The Analysis Framework, Asymptotic Notations and Basic Efficiency Classes. **Stacks**: What is a Stack? How Stacks are used, Stack ADT, Applications, Implementation, Comparison of Implementations. **Queues**: What is a Queue? How are Queues Used? Queue ADT, Applications, Implementation **Linked Lists:** What is a Linked List? Why Linked Lists?, Arrays Overview, Comparison of Linked Lists with Arrays & Dynamic Arrays, Singly Linked Lists, Doubly Linked Lists, Circular Linked Lists.	15
2	**Trees**: What is a Tree? Binary Trees, Types of Binary Trees, Properties of Binary Trees, Binary Tree Traversals, Generic Trees (N-ary Trees), Binary Search Trees (BSTs), Balanced Binary Search Trees, AVL (Adelson-Velskii and Landis) Trees. **Graph Algorithms:** Introduction, Applications of Graphs, Graph Representation, Graph Traversals, Topological Sort, Shortest Path Algorithms, Minimal Spanning Tree. **Sorting**: What is Sorting? Why is Sorting Necessary? Classification of Sorting Algorithms, Bubble Sort, Selection Sort, Insertion Sort, Merge Sort, Heap Sort, Quick Sort, Comparison of Sorting Algorithms. **Searching**: What is Searching? Why do we need Searching? Types of Searching, Unordered Linear Search, Sorted/Ordered Linear Search, Binary Search, Interpolation Search, Comparing Basic Searching Algorithms.	15
Case Study:		
1	A software company is designing an undo/redo feature for a text editor.	
2	An airport check-in system uses a queue to manage passengers waiting for boarding passes.	

Reference Books:

1. Introduction to Design and Analysis of Algorithms by Anany Levitin 3rd Ed Publisher: Pearson
2. Data Structures and Algorithms Made Easy by Narasimha Karumanchi Publisher: CareerMonk
3. Problem Solving with Algorithms and Data Structures Using Python by Bradley N. Miller & David L. Ranum Publisher: Franklin, Beedle & Associates Fundamentals
4. Python Algorithms (2nd Edition) by Magnus Lie Hetland Publisher: Apress
5. Open Data Structures in Python, https://opendatastructures.org/ods-python.pdf

Semester End Evaluation (50 Marks) **Time : 2 Hours**

Paper Pattern

Question No.	Questions	Total Marks : 50
Q1	Attempt 3 out of 5	15
Q2	Attempt 3 out of 5	15
Q3	Attempt 3 out of 5	15
Q4	Case Study	05

Practical Syllabus

Sr. No	List of Practical	No. of. Lectures	CO Mapping
1	Implement stack and demonstrate Push, Pop and Peek operations.	3	CO1
2	Queue using Array and Linked List: Implement a Queue ADT	3	CO1, CO2
3	Singly Linked List Operations: Write a program to implement a Singly Linked List with the following operations: ● Insert at beginning ● Insert at end ● Insert at given position ● Delete from beginning ● Delete from end ● Search an element ● Display list	3	CO2
4	Binary Tree Traversals: Write a program to create a Binary Tree. Implement Preorder, Inorder, and Postorder Traversals.	3	CO3
5	Graph Representation and Traversals: a.Depth First Search (DFS) b.Breadth First Search (BFS)	3	CO3
6	Shortest Path using Dijkstra's Algorithm: Implement Dijkstra's Algorithm to find the shortest path from a source node to all other nodes in a weighted graph.	3	CO3
7	Write a program to compute MST (Minimum Spanning Tree) for a connected graph using Prim's Algorithm	3	CO3
8	Implementing and Analysing Sorting Algorithms: a. Bubble Sort b.Insertion Sort c. Selection Sort	3	CO4
9	Sorting Algorithm Performance Comparison: a.Merge Sort b.Quick Sort	3	CO4
10	Searching Techniques Comparison Implement: a.Linear Search b.Binary Search	3	CO4

Semester End Practical Evaluation **Time: 2 Hours**

Question No.	Questions	Total Marks
Q.1	Program	30
Q.2	Journal	10
Q.3	Viva & Attendance	10

BOS	**Mathematics, Statistics and Computer Application**				
Course	**Data Warehousing**				
Course Code	**HUSDS210**	**Level**	**5.0**		
		Type	**Theory**	**Practical**	**Total**
Semester	**III**	**Credits**	2	1	3
Type	**Major**	**No of Teaching Hours**	30	30	60

Evaluation Pattern	**Total Marks**	**Semester End**	**Continuous**	**Practical**
	150	50	50	50

Learning Objectives	
1	Understand the fundamentals of data warehousing and business intelligence and their role in strategic decision-making systems.
2	Describe modern data warehouse architectures, including enterprise warehouses, data marts, and cloud-based warehousing platforms.
3	Develop knowledge of data integration techniques, including ETL/ELT processes and automated data pipeline concepts.
4	Apply dimensional modeling techniques such as star schema, snowflake schema, and advanced warehouse design concepts.
5	Explore current industry practices such as real-time data warehousing, governance, and BI tool integration for analytics applications.

Course Outcomes	
After successful completion of this course, students would be able to: -	
CO1	Explain the importance of data warehousing and business intelligence in decision-making, and describe data warehouse architecture, components, and modern cloud warehousing approaches.
CO2	Apply ETL/ELT concepts and demonstrate automated data pipeline processes.
CO3	Analyze and differentiate dimensional models using star and snowflake schemas, including slowly changing dimensions.
CO4	Apply OLAP techniques and use BI tools for business reporting and generating insights.

Modules at Glance

Module No.	**Content**	**No. of Hours**	**CO Mapping**
1	Foundations and Modern Data Warehousing Architecture	15	CO1, CO2
2	Data Processing, Modeling, and Modern Industry Practices	15	CO3, CO4

Syllabus

Module No.	Content	No. of Lectures
1	**Foundations and Modern Data Warehousing Architecture** **Evolution of Decision Support and Business Intelligence** Need for strategic information systems, Operational databases vs Analytical systems, Data Warehouse definition and characteristics, Evolution from DSS to modern Business Intelligence and Data Analytics platforms. **Data Warehouse Concepts and Components** Need for data warehousing, Features and properties of a Data Warehouse, Data Warehouse components, Metadata repository, Data marts and enterprise data warehouses. **Modern Data Warehouse Architecture** Three-tier architecture, Centralized and distributed warehouses, Cloud-based Data Warehousing concepts, Introduction to modern cloud platforms such as Snowflake, Amazon Redshift, Google BigQuery, Azure Synapse. **Data Lake, Data Warehouse, and Lakehouse** Difference between Data Warehouse and Data Lake, Introduction to Lakehouse architecture (Databricks concept), Role of warehousing in Big Data ecosystems. **Data Governance and Data Quality (Industry Requirement)** Data consistency, accuracy, data lineage, compliance needs, Introduction to data catalog and governance tools.	15
2	**Data Processing, Modeling, and Modern Industry Practices** **Data Warehouse Processing and Integration** ETL in Data Warehousing, ETL vs ELT, Modern ELT pipelines, Data transformation using tools like dbt, Data pipeline orchestration concepts using Apache Airflow. **OLAP and Analytical Processing** Introduction to OLAP, Characteristics of OLAP systems, OLTP vs OLAP, OLAP operations (roll-up, drill-down, slice, dice, pivot), Types of OLAP (MOLAP, ROLAP, HOLAP). **Real-Time Data Warehousing and Streaming Analytics** Need for real-time analytics, Introduction to streaming data integration using Apache Kafka, Spark Streaming, and cloud streaming services. **Data Warehouse Modeling and Dimensional Design** Conceptual modeling, Dimensional modeling concepts, Fact and Dimension tables, Measures and granularity, Star schema, Snowflake schema, Fact constellation. **Advanced Dimensional Concepts (Industry Standard)** Slowly Changing Dimensions (SCD Types 1, 2, 3), Types of fact tables (transaction, snapshot, accumulating snapshot). **Business Intelligence and Applications** Integration of data warehouses with BI tools, Introduction to Power BI/Tableau reporting, Warehousing support for AI/ML-driven analytics and decision-making.	15

Case Study Scenario	
M1	**Retail Business Intelligence Implementation** A retail company faced difficulty generating monthly sales and customer trend reports using its OLTP system. It implemented a Data Warehouse with star schema (Sales Fact and Product, Customer, Time Dimensions). The warehouse was connected to Microsoft Power BI for dashboards. This improved reporting speed and decision-making. **Concepts Covered:** OLTP vs OLAP, DW characteristics, Star schema, BI integration.
M2	**Data Governance in Banking** A bank faced data inconsistency across departments. It implemented ETL validation rules, metadata repository, and data lineage tracking. This improved data accuracy and regulatory compliance. **Concepts Covered:** Data governance, Data quality, Metadata, ETL.

Reference Books:

1. Fundamentals of Computer Graphics, Steve Marschner, Peter Shirley, CRC Press, 4th, 2016
2. The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling, Ralph Kimball, Margy Ross, John Wiley & Sons, 3rd, 2013
3. Data Mining: Concepts and Techniques, Jiawei Han, Micheline Kamber, Jian Pei, Morgan Kaufmann (Elsevier), 3rd, 2011
4. Mastering Data Warehouse Design: Relational and Dimensional Techniques, Claudia Imhoff, Nicholas Galemmo, Jonathan G. Geiger, John Wiley & Sons, 1st, 2003
5. Data Warehousing for Dummies, Alan Simon, Wiley Publishing, 1st, 2010

Semester End Evaluation (50 Marks) **Time : 2 Hours**

Paper Pattern

Question No.	Questions	Total Marks : 50
Q1	Attempt 3 out of 5	15
Q2	Attempt 3 out of 5	15
Q3	Attempt 3 out of 5	15
Q4	Case Study	05

Practical Syllabus

Sr. No	List of Practical	No. of. Lectures	CO Mapping
1	**Implementation of Star and Snowflake Schemas** **Objective:** Create a physical database structure based on a conceptual model. Write SQL DDL to create **Fact** and **Dimension** tables. Establish primary and foreign key relationships that enforce referential integrity within a Star or Snowflake schema.	3	CO3
2	**Complex Joins for De-normalization** **Objective:** Flatten normalized data into a warehouse-ready format. Use **Self-Joins**, **Multiple Inner/Outer Joins**, and **Cross Joins** to combine data from 5+ normalized tables into a single wide "denormalized" view for reporting.	3	CO2
3	**Advanced Aggregations with ROLLUP and CUBE** **Objective:** Create multi-level summary reports. Use the GROUP BY ROLLUP and GROUP BY CUBE clauses to generate hierarchical subtotals (e.g., Sales by Day > Month > Year) in a single query result set.	3	CO4
4	**Ranking and Window Functions** Analyze data relative to other rows without grouping. Implement RANK(), DENSE_RANK(), and ROW_NUMBER() to find "Top N" products per category or identify the highest-earning employees in each department.	3	CO4
5	**Time-Series Analysis with LEAD and LAG** **Objective:** Compare current performance against previous periods. Use LAG() and LEAD() window functions to calculate **Month-over-Month (MoM) growth** or identify trends in historical sales data.	3	CO4
6	**Working with Common Table Expressions (CTEs)** **Objective:** Simplify complex, nested logic for readability. Rewrite deep subqueries into named **CTEs** (using the WITH clause). Practice **Recursive CTEs** to traverse hierarchical data like an organizational chart (Employee -> Manager).	3	CO2
7	**Pivot and Unpivot Operations** **Objective:** Transform row-oriented data into column-oriented reports. Use the PIVOT operator (or CASE WHEN logic) to turn monthly sales rows into columns for a "Side-by-Side" yearly comparison report.	3	CO4
8	**Handling "Slowly Changing Dimensions" (SCD)** **Objective:** Manage historical changes in dimension data. Write an UPDATE/INSERT script to implement **SCD Type 2**. This involves using SQL to expire old records (setting an end_date) and inserting new versions of a record to keep history.	3	CO3

9	**Materialized Views and Indexing for Performance** **Objective:** Optimize query speed for large datasets. Create **Materialized Views** to pre-calculate heavy aggregations. Compare the execution plan (using EXPLAIN) of a query before and after adding **B-Tree** or **Bitmap indexes**.	3	CO1
10	**Data Quality Checks and Constraints** **Objective:** Ensure "GIGO" (Garbage In, Garbage Out) doesn't happen. Write SQL queries to identify "dirty data" (nulls, duplicates, or outliers) and use CHECK constraints or Triggers to prevent inconsistent data from entering the warehouse.	3	CO2

Semester End Practical Evaluation **Time: 2 Hours**

Question No.	Questions	Total Marks
Q.1	Program	30
Q.2	Journal	10
Q.3	Viva & Attendance	10

BOS	**Mathematics, Statistics & Computer Application**				
Course	**Data Mining**				
Course Code	**HUSDS209**	**Level**	**5.0**		
			Theory	**Practical**	**Total**
Semester	**III**	**Credit**	02	01	03
Type	**Major**	**No of Teaching hours**	30	30	60

Evaluation/ Assessment	**Total Marks**	**Semester End**	**Continuous**	**Practical**
	150	50	50	50

Learning Objectives	
1	To learn concept of Data Mining
2	To know data mining issues and its social implications.
3	To understand Data Mining Preprocessing.
4	To know the concept of Classification.
5	To learn Clustering and Prediction.
6	To understand Graph Mining, Social Network Analysis, and Multi-relational Data Mining

Course Outcomes	
CO1	Ability to acquire knowledge of Data Mining.
CO2	Ability to gain information about Data mining issues and its social implication.
CO3	Proficiency in preprocessing data.
CO4	Acquire knowledge of various classification methods.
CO5	To attain various techniques of clustering and prediction.
CO6	Ability to gain knowledge on Graph mining, Social Network Analysis and Multi Relational Data Mining.

Modules At GlanceSyllabus

Module No.	**Content**	**No. of Hours**	**Mapping with CO**
1	Data Mining Foundations & Preprocessing	15	CO1,CO2, CO3
2	Classification, Clustering & Advanced Mining	15	CO4, CO5,CO6

Syllabus

Module No.	Content	No. of Lectures
1	**Data Mining Foundations & Preprocessing** **1. Introduction to Data Mining** Evolution of Database Systems, Data Mining vs KDD, CRISP-DM Lifecycle (Industry Standard), Applications in Banking, Healthcare, E-commerce, Data Mining Issues & Social Implications, Ethics, Privacy & Responsible AI **2. Data Repositories** Data Warehouse, Data Lakes, Big Data Ecosystem (Introductory Overview) **3. Data Preprocessing** Data Cleaning, Missing Value Treatment, Outlier Detection,Data Integration & Transformation, Data Reduction Techniques, Feature Engineering (Industry Addition – 20% enhancement),Tools: Python (Pandas, NumPy – conceptual introduction) **4.Association Rule Mining** Frequent Itemsets, Apriori Algorithm (Conceptual),Market Basket Analysis (Case Study based)	15
2	**Classification, Clustering & Advanced Mining** **1. Classification & Prediction** What is Classification?, Decision Tree, Naïve Bayes, k-NN, Logistic Regression (Industry Addition), Model Evaluation: Accuracy, Precision, Recall, Confusion Matrix **2. Cluster Analysis** Types of Data in Clustering, k-Means Algorithm,Hierarchical Clustering, DBSCAN (Industry-relevant Density-based method), Use Case: Customer Segmentation **3. Graph & Social Network Mining** Introduction to Graph Mining, Centrality Measures (basic), and Social Network Analysis Applications **4. Text & Web Mining** Introduction to Text Mining, TF-IDF (basic idea), Web Mining Overview Real-world Application: Sentiment Analysis (Conceptual)	15
Case Study Scenarios		
M1	A retail supermarket chain operating across India has launched a loyalty card program. Over the past year, it has collected transaction data from 50,000 customers.The data contains:Missing product IDs,Duplicate transaction entries Inconsistent naming of products (e.g., “Milk 1L”, “1 Litre Milk”, “Milk-1L”) Large number of rarely purchased items The management wants to analyze customer buying patterns to improve product placement and increase bundled sales.Design a complete data preprocessing plan before applying any mining technique. Clearly explain the steps you would apply and justify why each step is necessary.	

M2	A telecom company is facing a high customer churn rate. The company has collected the following data: ● Customer age ● Monthly bill amount ● Contract duration ● Number of complaints ● Internet usage pattern ● Payment method ● Churn status (Yes/No) Management wants to: ● Predict which customers are likely to leave ● Group customers based on usage behavior ● Design retention strategies Apply clustering concepts to segment customers based on their usage patterns. Explain how these segments can help the company make strategic business decisions.	

Reference Books:

1.Data Mining: Introductory and Advanced Topics,M. H. Dunham ,Pearson 2nd Edition

2.Data Mining: Concepts and TechniquesJ. Han and M. Elsevier ,2nd Edition

3“Data Mining” A Knowledge Discovery Approach Krzysztof J, Cios,W. Pedrycz, R, W.Świniarski, L.A. Kurgan Springer 2nd

4.Data Mining Concepts & Techniques,Han & Kamber TMH ,2nd Edition

Semester End Evaluation (50 Marks) **Time: 2 Hr**

Paper Pattern

Question No	Questions	Total Marks: 50
Q1	Attempt any 3 out of 5	**15**
Q2	Attempt any 3 out of 5	**15**
Q3	Attempt any 3 out of 5	**15**
Q4	Case Study	**05**

Practical Syllabus

Sr No.	List of Practical	No. of Lectures	CO Mapping
1	**Introduction to WEKA Interface** Install WEKA and explore its components: Explorer, Experimenter, Knowledge Flow. Load sample datasets and understand ARFF/CSV format.	3	CO1
2	**Data Import and Dataset Understanding** Import datasets (CSV/ARFF) in WEKA, examine attributes, summary statistics, and visualize data using preprocessing tools.	3	CO1,CO2
3	**Data Preprocessing Techniques** Perform preprocessing tasks such as handling missing values, normalization, discretization, and filtering attributes using WEKA filters.	3	CO3
4	**Data Cleaning and Transformation** Apply attribute selection, remove noisy data, transform datasets using filters such as Remove, ReplaceMissingValues, and Normalize.	3	CO3
5	**Association Rule Mining using Apriori** Apply the Apriori algorithm to generate association rules. Analyze support, confidence, and lift values. Perform a Market Basket Analysis case study.	3	CO4
6	**Classification using Decision Tree** Implement Decision Tree (J48) classification in WEKA. Train and test the model using different datasets and analyze results.	3	CO4
7	**Classification using Naïve Bayes and k-NN** Implement Naïve Bayes and IBk (k-NN) algorithms. Compare performance using evaluation metrics.	3	CO4
8	**Model Evaluation Techniques** Evaluate classification models using accuracy, precision, recall, F-measure, and confusion matrix with cross-validation.	3	CO4,CO5
9	**Clustering using K-Means** Perform clustering using the SimpleKMeans algorithm and analyze cluster formation with visualization tools.	3	CO5
10	**Density-Based Clustering (DBSCAN) and Use Case** Implement density-based clustering in WEKA and analyze customer segmentation datasets.	3	CO5,CO6

Semester End Practical Evaluation **Time: 2 Hours**

Question No.	Questions	Total Marks
Q.1	Practical Questions	40
Q.2	Journal	05
Q.3	Viva & Attendance	05

BOS	**Mathematics, Statistics and Computer Application**				
Course	**Probability Theory with Discrete Distributions**				
Course Code	**HUSDS211**	**Level**	**5.0**		
		Type	**Theory**	**Practical**	**Total**
Semester	**III**	**Credits**	2	1	3
Type	**Minor**	**No of Teaching Hours**	30	30	60
Evaluation Pattern	**Total Marks**	**Semester End**	**Continuous**	**Practical**	
	150	50	50	50	

Learning Objectives	
1	To introduce the fundamental concepts of probability theory, including random experiments, sample spaces, events, and probability models used to analyze uncertainty in real-life situations.
2	To develop an understanding of counting principles, permutations, and combinations for solving probability problems and constructing probability models.
3	To explain conditional probability, multiplication theorem, and Bayes' theorem, enabling students to analyze dependent events and posterior probabilities in practical applications.
4	To introduce the concepts of sensitivity and specificity of a procedure and demonstrate the application of Bayes' theorem in analyzing false positive and false negative outcomes in diagnostic and decision-making processes.
5	To develop an understanding of random variables and standard discrete probability distributions, including probability mass function, distribution function, expectation, and variance, and apply these concepts to solve real-life numerical problems.

Course Outcomes	
After successful completion of this course, students would be able to: -	
CO1	Students will be able to understand and apply fundamental concepts of probability.
CO2	Students will be able to analyze and solve probability problems using conditional probability, independence of events, and important probability theorems such as addition theorem, multiplication theorem, and Bayes' theorem with appropriate proofs and applications.
CO3	Students will be able to understand and apply the concepts of discrete random variables and distribution functions.
CO4	Students will be able to analyze and solve problems involving standard discrete probability distributions such as Discrete Uniform, Bernoulli, Binomial, Poisson, Hypergeometric, and Geometric distributions, including derivation and application of their mean and variance.

Modules at Glance

Module No.	Content	No. of Hours	CO Mapping
1	Elementary Probability Theory	15	CO1, CO2
2	Discrete Random Variables and Standard Discrete Probability Distributions	15	CO3, CO4

Syllabus

Module No.	Content	No. of Lectures
1	**Elementary Probability Theory** **Definitions:** Trial, random experiment, sample point and sample space. Definition of an event and different types of events: compound event, complementary event, equally likely events, certain event, impossible event, mutually exclusive and exhaustive events. **Different definitions of Probability:** Classical (Mathematical), Empirical(Statistical) and Axiomatic definitions of Probability. Properties of probability Conditional probability. Independence of events, pairwise and mutual independence of three events. **Theorems (with proof)and their applications:** i. Addition theorem on probability for two and three events ii. Multiplication theorem on probability for two events. iii. Bayes' theorem.	15
2	**Discrete Random Variables and Standard Discrete Probability Distributions:** **Random Variables and Distribution Functions**:Definition of a random variable.Types of random variables: discrete and continuous random variables.Probability Mass Function (p.m.f.) – definition and properties. Cumulative Distribution Function (c.d.f.) and its properties**.** **Moments and Measures of Shape:**Raw moments and central moments (definitions only) and their relationships up to order four. Concepts of skewness and kurtosis and their interpretation for random variables. **Expectation and Variance**:Mathematical expectation and variance of a random variable. Important theorems and properties of expectation and variance with proofs. **Joint Distributions:**Joint probability mass function of two discrete random variables.Marginal and conditional distributions. Standard Discrete Probability Distributions:Definition and derivation of mean and variance for the following distributions: i) Discrete Uniform Distribution ii) Bernoulli Distribution iii) Binomial Distribution iv) Poisson Distribution v) Hypergeometric Distribution vi) Geometric Distribution	15
Case Study Scenario		
M1	A university is analyzing participation of students in two extracurricular activities: sports and music club.In a college survey of 200 students:80 students participate in sports,70 students participate in music club and 30 students participate in both sports and music.The administration wants to find the probability that a randomly selected student participates in at least one of the two activities.	

M2	A hospital emergency department studies the number of patients arriving per hour. Past records show that on average 3 patients arrive per hour. The arrivals are independent and occur randomly over time.Find the probability that exactly two patients arrive in an hour.

Reference Books:

1. Statistical Methods, G.W. Snedecor, W.G. Cochran, John Wiley & sons, 1991, Eighth Edition
2. Fundamentals of Applied Statistics, Gupta and Kapoor, S.Chand and Sons, New Delhi, 2014, Fourth Edition
3. An Introductory Statistics, Kennedy and Gentle.
4. Modern Elementary Statistics, Freund J.E., Pearson Publication, 2006, Twelfth Edition.
5. Probability, Statistics, Design of Experiments and Queuing theory with applications Computer Science, Trivedi K.S., Prentice Hall of India, New Delhi,2001, Second Edition.
6. A First course in Probability, Ross, Pearson Publication, 2013, Ninth Edition.
7. A First Course in Probability and Statistics, L. S. Prakasa Rao, World Scientific Publishing Co Pte Ltd, 2008.
8. Applied Probability Models, D. L. Minh, Brooks/Cole, 2000.

Semester End Evaluation (50 Marks) **Time : 2 Hours**

Paper Pattern

Question No.	Questions	Total Marks : 50
Q1	Attempt 3 out of 5	15
Q2	Attempt 3 out of 5	15
Q3	Attempt 3 out of 5	15
Q4	Case Study	05

Practical Syllabus

Sr. No	List of Practical (Conducted in Python)	No. of. Lectures	CO Mapping
1	To understand the concepts of trial, random experiment, sample point, and sample space using Python.	3	CO1
2	To illustrate different types of events using Python.	3	CO1
3	Compute probability theoretically and compare with experimental probability.	3	CO1
4	To compute conditional probability and verify independence of events using Python.	3	CO2
5	To verify important probability theorems using Python.	3	CO2
6	To understand the concept of a random variable and probability mass function (p.m.f.) using Python.	3	CO3
7	To study the cumulative distribution function (c.d.f.) and its properties.	3	CO3
8	To compute raw moments, central moments, skewness, and kurtosis for a discrete random variable	3	CO4
9	To calculate expectation, variance, and joint probability mass functions using Python.	3	CO4
10	To study standard discrete probability distributions using Python.	3	CO4

Semester End Practical Evaluation **Time: 2 Hours**

Question No.	Questions	Total Marks
Q.1	Program/Problems	30
Q.2	Journal	10
Q.3	Viva & Attendance	10

BOS	**Mathematics, Statistics and Computer Application**				
Course	**Scala for Data Science**				
Course Code	**HUSDS212**	**Level**	**5.0**		
		Type	**Theory**	**Practical**	**Total**
Semester	**III**	**Credits**	2	1	3
Type	**SEC**	**No of Teaching Hours**	30	30	60
Evaluation Pattern	**Total Marks**	**Semester End**	**Continuous**		**Practical**
	150	50	50		50

Learning Objectives	
1	To provide students with a strong foundation in the basics of Scala, including syntax, data types, control structures, and functions.
2	To provide students with an understanding of statistical computations and data manipulation techniques, using Scala collections and basic programming constructs.
3	To provide students with the ability to perform numerical and matrix operations using libraries such as Breeze.
4	To provide students with exposure to data analysis and introductory machine learning concepts, along with basic data processing using Apache Spark.

Course Outcomes	
CO1	Students will have the ability to understand the fundamentals of Scala programming, including syntax, data types, control structures, and functions.
CO2	Students will have the ability to perform statistical computations and manipulate data using Scala collections and basic programming techniques.
CO3	Students will have the ability to apply numerical computing concepts using libraries such as Breeze for vector and matrix operations.
CO4	Students will have the ability to analyze datasets and implement basic data processing and machine learning tasks using Scala and introductory concepts of Apache Spark.

Modules At Glance

Module No.	**Content**	**No. of Hours**	**Mapping with CO**
1	Foundations of Scala for Data Science	15	CO1, CO2
2	Data Analysis and Machine Learning using Scala	15	CO3, CO4, CO5

Syllabus

Module No.	Content	No. of Hours
1	1. Introduction to Scala and Development Environment: Overview and features of Scala, Applications of Scala in Data Science,Installing Scala and SBT, Writing and executing Scala programs 2. Scala Programming Basics: Variables and data types, Scala collections (List, Array, Vector), Input and output operations, Basic operations on collections 3. Statistical Computations using Scala: Mean, median, and mode, Variance and standard deviation, Frequency distribution, Correlation concepts. 4. Introduction to Breeze Library: Overview of Breeze library, Dense vectors and matrices, Matrix operations (transpose, determinant), Element-wise matrix operations 5. Data Handling in Scala: Reading data from CSV files, Data filtering and sorting, Handling missing values, Word frequency analysis in text data	15
2	1. Data Visualization: Introduction to data visualization, Scatter plots, Histograms, Line graphs and combined plots 2. Time Series and Statistical Analysis: Moving averages, Time series data handling, Trend analysis, Data summarization techniques 3. Machine Learning Basics in Scala: Introduction to machine learning, Linear regression, Logistic regression, Model prediction concepts 4. Clustering and Similarity Measures: Distance measures (Euclidean distance), Nearest neighbor concept, K-means clustering, Applications in data science 5. Introduction to Big Data Processing with Spark: Introduction to Apache Spark, Spark DataFrames, Data filtering and group operations, Joining datasets and simple ML pipelines	15
Case Study Scenario		
M1	A retail company has collected daily sales data in a CSV file containing fields such as product name, quantity sold, and price. However, the dataset contains missing values and inconsistent entries. The company wants to analyze this data using Scala to gain insights into sales performance. **Question:** Using Scala (with collections and basic data handling techniques), write a program to read the CSV data, clean missing values, and compute key statistical measures (mean, median, and standard deviation of sales). Additionally, perform sorting and display the top-performing products based on total sales.	

M2	A startup is analyzing customer data to understand purchasing behavior. The dataset includes features such as customer age, income, and spending score. The company wants to visualize trends and group similar customers for targeted marketing. **Question:** Using Scala, implement data visualization (scatter plot or histogram) and apply a clustering technique (such as K-means) to group customers based on their features. Explain how the clusters can help in making business decisions.

Reference Books:

1.Programming in Scala, Odersky M., Spoon L., Venners B., Programming in Scala, Artima Press.
2.Programming Scala, Wampler D., Payne A., Programming Scala, O'Reilly Media.
3.Hands-On Scala Programming, Haoyi L., Hands-On Scala Programming, Li Haoyi Publications.
4.Scala Cookbook, Alexander A., Scala Cookbook, O'Reilly Media.
5.Hands-On Scala Programming, Haoyi L., Hands-On Scala Programming, Li Haoyi Publications.

Semester End Evaluation (50 Marks) **Time: 2 Hr**

Paper Pattern

Question No	Questions	Total Marks: 50
Q1	Attempt any 3 out of 5	**15**
Q2	Attempt any 3 out of 5	**15**
Q3	Attempt any 3 out of 5	**15**
Q4	Case Study	**05**

Practical Syllabus

Sr No.	List of Practicals	No. of Lecture	CO Mapping
1	**Installation and Setup of Scala** Install Scala and SBT environment and write a simple Scala program to display output.	3	CO1
2	**Basic Scala Programming** Write programs using **variables, data types, and operators** to perform simple arithmetic operations.	3	CO1
3	**Statistical Computation** Write a Scala program to calculate **mean, median, and mode** for a given dataset.	3	CO2
4	**Variance and Standard Deviation** Write a Scala program to compute **variance and standard deviation** for a list of numbers.	3	CO2
5	**Breeze Vector Operations** Create **Breeze vectors** and perform operations such as **sum, mean, and dot product**.	3	CO3
6	**Breeze Matrix Operations** Create **Breeze matrices** and perform operations such as **transpose and determinant**.	3	CO3
7	**CSV Data Handling** Write a Scala program to **read a CSV dataset and compute basic statistics**.	3	CO2, CO4
8	**Data Visualization** Generate a **scatter plot or histogram** for a dataset using Scala visualization libraries.	3	CO4
9	**Linear Regression Model** Implement a **simple linear regression model** using Scala/Breeze.	3	CO3, CO4
10	**Apache Spark Data Processing** Use **Apache Spark with Scala** to perform **word count or dataset filtering operations**.	3	CO4

Semester End Practical Evaluation **Time: 2 Hours**

Question No.	Questions	Total Marks
Q.1	Practical Questions	40
Q.2	Journal	05
Q.3	Viva & Attendance	05