MANONMANIAM SUNDARANAR UNIVERSITY
TIRUNELVELI-627 012, TAMILNADU, INDIA

CENTRE FOR INFORMATION TECHNOLOGY AND ENGINEERING

Board of Studies Meeting Held on 29.03.2022

M.Sc. Data Analytics

(CBCS-University Department)

Regulations, Scheme and LOCF based Syllabus
For those who joined from the academic year 2022-2023 onwards

Submitted by
Chairman, BOS and Head,
Centre for Information Technology and Engineering,

to

The Registrar
Manonmaniam Sundaranar University
Tirunelveli - 12

MANONMANIAM SUNDARANAR UNIVERSITY
TIRUNELVELI-627 012, TAMILNADU, INDIA

Centre for Information Technology and Engineering
M.Sc. Data Analytics
(CBCS-University Department)

Regulations, Scheme and LOCF based Syllabus
For those who joined from the academic year 2022-2023 onwards

PREAMBLE

Vision of the Programme: M.Sc. degree programme in Data Analytics is designed to provide the students a long term career scope in the Data Analytics, Business Analytics and Artificial Intelligence industries of present and forthcoming decades. Through the programme, the students shall be exposed to the wide range of application possibilities of Data Analytics, Data storage and processing platforms, robust concepts and techniques for Data Analytics so that they acquire skills to develop real-time Data Analytics solutions to pursue career in contemporary IT industry and carry out research as time, environment and society dictates.

Curriculum Highlights: The M.Sc. degree programme in Data Analytics (M.Sc. DA) is a fine curriculum aimed squarely at producing graduates with multi-faceted skills needed to draw insights from complex and big data sets, and to be able to communicate those insights effectively. The Super-Child of five major fields, Mathematics, Statistics, Computer Science, Information Technology and Business Management, it is the product of the interdisciplinary group of experts and practicing professionals including mathematicians, statisticians, computer scientists, IT professionals, economists and operations researchers.

Uniqueness of M.Sc. DA programme: M.Sc. DA programme is an intensive 20-month learning experience designed to immerse students into the acquisition of practical knowledge and application of methods and techniques. The curriculum is carefully calibrated that would get updated continuously to meet the evolving challenges facing data scientists. The CITE department possesses classrooms, team rooms, study spaces,

and other amenities under one roof; with faculty members having experience of over two decades and industry expertise, the curriculum has contents capable of producing superior student outcomes.

PROGRAMME EDUCATIONAL OBJECTIVES (PEO)

The objectives of the programme are:

PEO1. To understand basic Data Analytics concepts and techniques for solutions in the IT industries and to carry out research as time, environment and society dictates.

PEO2. To remember basics concepts and techniques of Data Analytics.

PEO3. To acquire knowledge and skills in the statistical and mathematical concepts, programming techniques and computing tools for Data Analytics.

PEO4. To apply Data Analytics knowledge and skills in the wide range of Data Analytics based domains using Descriptive, Diagnostic, Predictive and Prescriptive approaches of Data Analytics.

PEO5. To acquire knowledge in the necessary ethics as a Data Analytics professional or Scientist.

PEO6. To understand the need for team work to solve complex problems in the wide range of Data Analytics based domains.

PROGRAMME SPECIFIC OUTCOMES (PSO)

PSO1. Knowledge: Able to remember basic Data Analytics concepts and techniques.

PSO2. Analytical Knowledge: Able to analyze and evaluate the data associated with different data analytics based problems to make insights.

PSO3. Analyze & Solve Complex Problems: Able to identify, plan, design and complete Data Analytics Projects.

PSO4. Research: Competent to solve complex problems using the statistical and mathematical concepts and programming techniques for Data Analytics.

PSO5. Modern Tool usage: Able to take up Data Analytics as time, environment and society dictates using modern computing tools and techniques.

PSO6. Ethics: Able to understand his/her ethical responsibility.

PSO7. Specific and Team work solving: Ability to understand each phase in the data analytics projects and work as an individual or a group to deliver best analysis and solutions.

PEO and PSO Mapping

	PSO1	PSO2	PSO3	PSO4	PSO5	PSO6	PSO7
PE01	S	S	S	M	M	L	M
PE02	S	S	S	S	M	L	M
PE03	S	S	S	S	S	L	L
PE04	S	S	S	S	S	L	L
PE05	L	L	M	L	L	S	M
PE06	L	L	L	L	L	M	S

* S- Strong, M – Middle, L – Low

A. REGULATIONS FOR M.Sc. DATA ANALYTICS

A1: Duration of the Programme:

The M.Sc. DA is a 2 years full time programme spread over four semesters.

A2: Eligibility for Admission:

The minimum eligibility conditions for admission to the M.Sc. DA programme are given below.

The candidates who seek admission into the first semester of the M.Sc. DA programme will be required to have obtained the Bachelor's / Master's degree or equivalent in Science (B.Sc. /M.Sc. /BCA /MCA) or Engineering (B.E. /M.E.) or Management (BBA /MBA) from Manonmaniam Sundaranar University or any other Indian University or equivalent in any one of the following disciplines:

1.Statistics

2.Mathematics

3. Business Administration

4.Information Technology

5.Information Technology and E-Commerce

6.Computer Science

7.Computer Applications

8. Any Engineering discipline

9. Any other discipline with Statistics/Mathematics/Computer Science/IT as a subject.

The minimum percentage of marks required for admission is based on the periodic regulations made by the University and the Government.

A3: Credit Requirement for the Degree:

The general Regulations of the Choice Based Credit System of Manonmaniam Sundaranar University are applicable to this programme. The University requirement for the M.Sc. programme is completion of 90 credits of course works. Out of 90 credits, 4 credits should be through the mini project, 10 credits should be through the 4^{th} semester major project work and remaining 76 credits should be through Core, Elective, Practical and Supportive Course papers. A Core course has 4 credits; elective, Supportive courses weigh 3 credits and Practical course weigh 2 credits. No candidate will be eligible for the Degree of M.Sc. DA unless the candidate has undergone the prescribed courses of study for a period not less than 4 semesters and has acquired 90 credits and other passing requirements in all courses of study. The marks, M_i obtained by the student in each course, i shall be multiplied by the credit of that course, C_i; such marks of all 'n' courses are added up and divided by the total credit (90) to obtain the Consolidated Percentage of Marks.

$$\text{Consolidated Percentage of Marks} = \frac{\sum_{i=1}^{n} C_i \times M_i}{\sum_{i=1}^{n} C_i}$$

A4: Attendance Requirement:

A candidate will be permitted to appear for the semester examination only if the candidate has not less than 75 percent attendance. The University condonation rules are

applicable for those who lack minimum of 75% attendance. The candidates with less than 60% attendance will have to repeat the concerned entire semester.

A5: Assessment

The assessment will comprise Continuous Internal Assessment (CIA) carrying a maximum of 25% marks and end-semester Examination carrying a maximum of 75% marks for each theory course (Core/Elective/Supportive Course). The Internal Assessment includes three Internal Tests, Assignment and Seminar. For practical courses, Mini Project and Major Project, the CIA is carried out for 50% marks and the external assessment is carried out for 50% marks. The external assessment includes University Practical Exam and Viva-Voce for Practical courses; Final Project Presentation, Project Report and Viva-Voce for Mini Project and Major Project.

Semester examination will be conducted for all courses of study, at the end of each Semester.

If a Student wants to carry out the final Major project work in 4^{th} semester in an IT company, the student can get permission from the concerned Project Supervisor, Faculty Members of that semester, Programme Coordinator and Head of the Department after submitting the Acceptance Letter from the IT Company. The student can also carry out industrial internship during the summer and winter vacations of the respective semesters during their course of study.

A6: Passing Requirements

A candidate who secures not less than 50 percent marks in end-semester examination and not less than 50 percent of the total marks (Continuous Internal Assessment + end-semester examination) in any course of study will be declared to have passed the course.

A Candidate who successfully completes the course and satisfies the passing requirements in all the courses of study and curricular requirements will be declared to have qualified for the award of the Degree.

A7: Classification of successful candidates

The candidates who have passed all theory courses, practical courses and Projects shall be classified as follows. Total Marks secured in theory courses, practical courses and Projects are used to obtain the Consolidated Percentage of Marks as mentioned in Regulation A3. The classifications are as follows:

Marks Overall %	Classification
1. 75% and above with a First attempt Pass in all Courses	I Class with Distinction
2. i) 75% above from multiple attempts	I Class
ii) 60% to below 75%	I Class
3. 50% to below 60%	II Class

A8: Academic Session

The academic year normally begins in July every year and ends in April.

A9: Power to Modify

The University may from time to time revise, amend or change the regulations, scheme of examinations and syllabus, if found necessary and such amendments, changes shall come into effect from the date prescribed.

These regulations will come into effect from the academic year 2022-2023 onwards.

B.SCHEME FOR M.Sc. DATA ANALYTICS

(For those who joined from the academic year 2022-2023 onwards)

Sem-ester	Title of the Course	Category*	Hrs/ week	Credits	Maximum Marks			Passing Minimum Percentage	
					Inte-rnal	Exte-rnal	Total	Exter nal	Total
FIRST SEMESTER									
I	Data Analytics (E - Pathshala)	C	4	4	25	75	100	50	50
I	Mathematics for Data Analytics	C	4	4	25	75	100	50	50
I	Statistics for Data Analytics	C	4	4	25	75	100	50	50
I	SQL and NoSQL Database Management Systems	C	4	4	25	75	100	50	50
I	Data Visualization	C	4	4	25	75	100	50	50
I	SQL and NoSQL Database Management Systems Laboratory	L	2	2	50	50	100	50	50
I	Data Visualization Laboratory	L	2	2	50	50	100	50	50
I Semester Total Credits				**24**					
SECOND SEMESTER									
II	MOOCs Supportive Course I	S	3	3	25	75	100	50	50
II	Machine Learning	C	4	4	25	75	100	50	50
II	Descriptive and Discovery Analytics	C	4	4	25	75	100	50	50
II	Elective (Group A)	E	3	3	25	75	100	50	50
II	Elective (Group A)	E	3	3	25	75	100	50	50
II	Elective (Group A)	E	3	3	25	75	100	50	50
II	Machine Learning Laboratory	L	2	2	50	50	100	50	50
II	Descriptive and Discovery Analytics Laboratory	L	2	2	50	50	100	50	50
II Semester Total Credits				**24**					
THIRD SEMESTER									
III	MOOCs Supportive Course II	S	3	3	25	75	100	50	50
III	Predictive and Prescriptive Analytics	C	4	4	25	75	100	50	50

III	Ethics for Data Scientists	C	4	4	25	75	100	50	50
III	Elective (Group B)	E	3	3	25	75	100	50	50
III	Elective (Group B)	E	3	3	25	75	100	50	50
III	Predictive and Prescriptive Analytics Laboratory	L	2	2	50	50	100	50	50
III	Mini Project	P	4	4	50	50	100	50	50
III Semester Total Credits				**23**					
FOURTH SEMESTER									
IV	Elective (Group C) (E-Pathshala)	E	3	3	25	75	100	50	50
IV	Elective (Group C)	E	3	3	25	75	100	50	50
IV	Elective (Group C)	E	3	3	25	75	100	50	50
IV	Major Project	P	10	10	50	50	100	50	50
IV Semester Total Credits				**19**					
OVERALL TOTAL CREDITS				**90**					

***C-Core, L-Laboratory, S-Supportive, E-Elective, P-Project**

Electives Courses for Semester II (Group A)									
Sl. No.	**Title of the Course**	**Category***	**Hrs/week**	**Credits**	**Maximum Marks**			**Passing Minimum Percentage**	
					Inte-rnal	**Exte-rnal**	**Total**	**External**	**Total**
A1	Programming with Python for Data Analytics	E	3	3	25	75	100	50	50
A2	Natural Language Processing	E	3	3	25	75	100	50	50
A3	Computer Vision and Applications	E	3	3	25	75	100	50	50
A4	Programming with Mangodb for Data Analytics	E	3	3	25	75	100	50	50
A5	Big Data Analytics on Genomic Data	E	3	3	25	75	100	50	50
Electives Courses for Semester III (Group B)									
B1	Big Data Security	E	3	3	25	75	100	50	50

B2	Deep Learning	E	3	3	25	75	100	50	50
B3	Data Mining Essentials for Data Analytics	E	3	3	25	75	100	50	50
B4	Data Visualization using Tableau	E	3	3	25	75	100	50	50
B5	Programming with Cassandra	E	3	3	25	75	100	50	50
Electives Courses for Semester IV (Group C)									
C1	Cloud Computing [E-PG Pathshala]	E	3	3	25	75	100	50	50
C2	HADOOP for Data Analytics	E	3	3	25	75	100	50	50
C3	Cloud Platforms in Industry	E	3	3	25	75	100	50	50
C4	Storm for Data Analytics	E	3	3	25	75	100	50	50
C5	Spark for Data Analytics	E	3	3	25	75	100	50	50

***C-Core, L-Laboratory, S-Supportive, E-Elective, P-Project**

MANONMANIAM SUNDARANAR UNIVERSITY

TIRUNELVELI, TAMILNADU

M.Sc DATA ANALYTICS DEGREE PROGRAMME

LIST OF CORES

(For those who joined from the academic year 2022-2023 onwards)

Sl. No.	Course Code	Semester	Course Name
1.		I	Data Analytics (E-Pathashala)
2.		I	Mathematics for Data Analytics
3.		I	Statistics for Data Analytics
4.		I	SQL and NoSQL Database Management Systems
5.		I	Data Visualization
6.		I	SQL and NoSQL Database Management Systems Laboratory
7.		I	Data Visualization Laboratory
8.		II	Machine Learning
9.		II	Descriptive and Discovery Analytics
10.		II	Machine Learning Laboratory
11.		II	Descriptive and Discovery Analytics Laboratory
12.		III	Predictive and Prescriptive Analytics
13.		III	Ethics for Data Scientists
14.		III	Predictive and Prescriptive Analytics Laboratory
15.		III	Mini Project
16.		IV	Major Project

MANONMANIAM SUNDARANAR UNIVERSITY

TIRUNELVELI, TAMILNADU

M.Sc DATA ANALYTICS DEGREE PROGRAMME

LIST OF ELECTIVES

(For those who joined from the academic year 2022-2023 onwards)

Sl. No.	Course Code	Semester	Course Name
Electives Courses for Semester II (Group A)			
1.		II	Programming with Python for Data Analytics
2.		II	Natural Language Processing
3.		II	Computer Vision and Applications
4.		II	Programming with Mangodb for Data Analytics
5.		II	Big Data Analytics on Genomic Data
Electives Courses for Semester III (Group B)			
6.		III	Big Data Security
7.		III	Deep Learning
8.		III	Data Mining Essentials for Data Analytics
9.		III	Data Visualization using Tableau
10.		III	Programming with Cassandra
Electives Courses for Semester IV (Group C)			
11.		IV	Cloud Computing [E-PG Pathshala]
12.		IV	HADOOP for Data Analytics
13.		IV	Cloud Platforms in Industry
14.		IV	Storm for Data Analytics
15.		IV	Spark for Data Analytics

C.SYLLABUS FOR M.Sc. DATA ANALYTICS

SEMESTER I

LIST OF COURSES

(For The Candidates Admitted From 2022-23 Onwards)

Sl. No.	Course Code	Course Name
1.		Data Analytics (E-Pathashala)
2.		Mathematics for Data Analytics
3.		Statistics for Data Analytics
4.		SQL and NoSQL Database Management Systems
5.		Data Visualization
6.		SQL and NoSQL Database Management Systems Laboratory
7.		Data Visualization Laboratory

Course Code	Course Name	**Category**	L	P	T	Credit
	DATA ANALYTICS (E-PATHSHALA)	C	4			4

Preamble

- To get introduced to the basic concepts in Data analytics.
- To understand the basic processing, storage and programming models for Data Analytics.
- To learn the different phases of Data Analytics Project.

Prerequisite

- Advanced Microsoft Excel Operations

Course Outcomes (COs)

On the successful completion of the course, students will be able to

Course Outcomes		Bloom's Level
CO1	Identify the scope of Data Analytics Project	Understand, Analyze
CO2	Define the basic hardware and programming requirements needed for a Data Analytics Project.	Remember, Apply
CO3	Plan the different phases of a Data Analytics Project	Apply, Create

Mapping with Programme Outcomes

COs	PSO1	PSO2	PSO3	PSO4	PSO5	PSO6	PSO7
CO1	S	S	M	M	M	L	L
CO2	M	S	S	M	S	L	M
CO3	M	M	M	M	M	L	S

Assessment Pattern

Bloom's Category	Continuous Internal Assessment (25)			Terminal Examination (75)
	I	II	III	
Remember	5	5	5	22
Understand	8	8	8	23
Apply	5	5	5	10
Analyze	5	5	5	10
Evaluate	0	0	0	0
Create	2	2	2	10

Syllabus

Unit I-Introduction to Analytics and Big Data: Data Analytics-Definition-importance--Big Data Analytics- Big Data approach to Analytics & Web– the actual big data-Evolution of Analytical Scalability-Analytic Processes and its evolution -Analytic Framework, Analytic Data Set (ADS) -Data Analysis -Problem Framing-Inference - Open Source Software**.**

(11 hrs)

Unit II-Data Analysis: Data Measurement-Types of data, Measurement Scales– Univariate analysis-Bivariate and Multivariate analysis- Probabilistic and Bayesian Approaches-Regression analysis. **(12 hrs)**

Unit III-Data Analytics-Mining Streams: Data stream - Data Stream processing model-Algorithms for streams-Filtering Data Streams-Moment-The Alon-Matias-Szegedy (AMS) Algorithm for Second Moments-Counting item sets-Sliding Windows-The Datar-Gionis-Indyk-Motwani Algorithm(DGIM)-Real Time application. **(14 hrs)**

Unit IV-Data Analytics –Association Rule Mining: Association Rule: Basic Concepts - Association rule discovery-The Apriori Algorithm-Mining Frequent Item sets: the Key Step-Computation Model for Finding Frequent Item sets. **(12 hrs)**

Unit V-Clustering: High Dimensional Data-Clustering High-Dimensional Data-Clustering Problem-Major Clustering Approaches-Introduction to Distance metrics-Non-Euclidean Case-Cohesion Metrics-Partitioning method **(11 hrs)**

Total: 60 hrs

Reference Books and URLs:

1. "A Guide to Big Data Analytics", Datameer.com.
2. "Big Data Analytics with R and Hadoop", Vignesh Prajapati, Packt Publications.
3. "Data Science and Big Data Analytics", D. Dietrich, B.Heller, B.Yang, EMC Education Services.
4. "Big Data Now", O'Reily Inc.
5. "A Comparison of Approaches to Large-Scale Data Analysis", DeWitt, S. Madden, and M. Stonebraker, SIGMOD Conference 2009.
6. www.researchgate.net/publication/273961581_Big_Data_analytics_with_applications
7. http://cs.ulb.ac.be/conferences
8. www.businessesgrow.com/2016/12/06/big-data-case-studies
9. https://epgp.inflibnet.ac.in/ahl.php?csrno=7
10. www.mercerindustries.com/wp-content/uploads/2015/02/Watson-Tutorial-Big-Data Business-Analytics

Course Code	Course Name	**Category**	L	P	T	Credit
	MATHEMATICS FOR DATA ANALYTICS	C	4			4

Preamble

- To understand the need of Mathematical concepts in the different stages of Data Analytics.
- To acquire knowledge in different Optimization and Linear Algebra concepts towards inference and prediction stages of Data Analytics.
- To acquire skills in a math processing tool.

Prerequisite

- Foundations of Mathematics

Course Outcomes

On the successful completion of the course, students will be able to

Course Outcomes		Bloom's Level
CO1	Define the basic optimization and linear algebra concepts for prediction and inference	Understand, Remember
CO2	Apply the right optimization technique at the Predictive stage of a Data Analytics project.	Remember, Apply
CO3	Acquire skills to program optimization and linear real world analytics problems in Octave.	Understand, Apply, Analyze, Evaluate, Create

Mapping with Programme Outcomes

COs	PSO1	PSO2	PSO3	PSO4	PSO5	PSO6	PSO7
CO1	S	S	M	M	L	L	L
CO2	S	S	S	S	M	L	L
CO3	S	S	S	M	S	L	M

Assessment Pattern

Bloom's Category	Continuous Internal Assessment (25)			Terminal Examination (75)
	I	II	III	
Remember	5	5	5	22
Understand	6	6	6	23
Apply	5	5	5	10
Analyze	5	5	5	10
Evaluate	2	2	2	5
Create	2	2	2	5

Syllabus

UNIT I Introduction: Machine Learning – Definition- Examples- Supervised Learning- Definition- Examples- Unsupervised Learning-Definition and Examples- Model Representation- Cost Function- Intuitions. **(12hrs)**

UNIT II Gradient Descent and Regularization: Gradient Descent- Intuitions- Gradient Descent for a Regression algorithm- Multiple Features- Gradient Descent on Multiple Features- Practice on Gradient Descent- Gradient Descent for Polynomial Regression- Normal Equation- Invertibility- Problem of Over fitting- Regularization- Cost function- Regularized Linear Regression. **(10hrs)**

UNIT III Linear Algebra: Matrix Representation- Examples of matrix Data- Vectors- examples- Representation- Matrix Addition- Scalar Multiplication- Matrix Multiplication properties- Matrix Vector Multiplication- Matrix Matrix Multiplication- Inverse and Transpose- Applications of Matrix operations on Real Time Data- Parallel Matrix Multiplication- Dimensionality Reduction by Principal Component Analysis and Eigen Values- Eigen Vectors. **(11hrs)**

UNIT IV Basic Operations of Octave: Octave Installation- Logical and Arithmetic Operations- Assignment of Different Variables- Exercises- Assigning Matrices- Vectors- Representation- Exercises- Histogram of Matrices- Diagonal Matrices- Help in Octave. **(13hrs)**

UNIT V Data Visualization and Processing using Octave: Finding the size of a Matrix- Loading Data into Octave- Viewing the Workspace of Octave- Accessing the elements of Matrix- Arithmetic operations on matrices: Addition, Multiplication, log, exponentiation, Transpose, Maximum and Minimum Value of a Matrix- Control Statements in Octave- Visualizing Data in Octave: Plotting Data, giving labels, axes and titles –Victimization- Vector implementation- Advantages- Assignments in Optimization. **(14hrs)**

Total: 60hrs

Reference Books and URLs:

1. Lectures of Professor Dr. Andrew Ng, Stanford University, Coursera.
2. "Matrix Computations", Gene H.Golub, Charles F.Van Loan, John Hopkins University Press.
3. "Eigen Values and Eigen Vectors in Data dimension Reduction for Regression", Randolf H. Reiss, B.S, San Marcos, Texas.
4. "Linear Algebra and its Applications", Gilbert Strang, Thomson Learning Inc., 4th Edition.
5. https://skymind.ai/wiki/eigenvector

Course Code	Course Name	**Category**	L	P	T	Credit
	STATISTICS FOR DATA ANALYTICS	C	4			4

Preamble

- To understand the need of Statistical concepts in the different stages of Data Analytics.
- To acquire statistical knowledge for different stages of Data Analytics.
- To acquire skills in using the right statistical concept at the right stage of Data Analytics.

Prerequisite

- Basic Statistics & Probability

Course Outcomes

On the successful completion of the course, students will be able to

Course Outcomes		Bloom's Level
CO1	Define the Descriptive, Inferential and Predictive statistical concepts for relevant stages of Data Analytics	Remember, Understand, Apply, Analyze
CO2	Acquire statistical knowledge for different stages of Data Analytics	Remember, Understand, Apply, Analyze
CO3	Apply the right statistical technique at right stage of a Data Analytics project	Apply, Analyze, Evaluate, Create

Mapping with Programme Outcomes

COs	PSO1	PSO2	PSO3	PSO4	PSO5	PSO6	PSO7
CO1	S	S	S	M	M	L	L
CO2	S	S	S	M	M	L	L
CO3	M	S	S	M	M	L	L

Assessment Pattern

Category	Continuous Internal Assessment (25)			Terminal Examination (75)
	I	II	III	
Remember	5	5	5	22
Understand	6	6	6	23
Apply	5	5	5	10
Analyze	5	5	5	10
Evaluate	2	2	2	5
Create	2	2	2	5

Syllabus

UNIT I Descriptive Statistics: Data - Describing Discrete Data - Continuous Data - Statistics for Discrete and Continuous Data- Outliers and Box Plots- Comparing Data Sets- Measures of Central Tendency- Measures of Scale- Relationship between Variables- Linear Model, Residual Analysis. **(12hrs)**

UNIT II Inferential Statistics: Examples of Prediction Problems - Probability- Determination of Probability-Examples - Conditional Probability- Random Variables- Parameters- Discrete Probability Models-Binomial Probability Model- Poisson Probability Model- Continuous Probability Models-Uniform Probability Model- Normal Distribution. **(10hrs)**

UNIT III Bayesian Statistics: Bayesian Statistics- Using Bayesian analysis to estimate a Proportion- Specifying a Prior for a Proportion- Calculating the Likelihood for a Proportion- Calculating the Posterior Distribution for a Proportion -Problems. **(11hrs)**

UNIT IV Concepts of Hypothesis Testing: Central Limit Theorem- Confidence Intervals- Testing Procedure- The Wilcoxon- Wilcoxon signed -rank Test -Wilcoxon rank-sum Test/ Mann Whitney U Test - Estimation and Confidence Interval based on Wilcoxon. **(13hrs)**

UNIT V Design of Experiments and Regression: Completely Randomized Designs- Randomized Pair Designs- Regression Experimental Design- Example- Observational Studies- Linear Regression and Applications. **(14hrs)**

Total: 60hrs

Reference Books and URLs:

1. "Statistics and Data Analytics", A.Abebe, J.Daniels, J.W.Macean, Statistical Computation Lab, Western Michigan University, Kalamazoo.
2. "A little Book for R using Bayesian Statistics", Avril Coghlan, Wellcome Trust Sanger Institute, U.K.
3. "Bayesian Statistics" (product code M249/04) by the Open University, available from the Open University Shop.
4. "Data Mining, Inference and Statistics", Trevor Hasti, Robert Tibshirani, Jerome Friedman, 2nd Edition, Springer Series in Statistics.
5. " Bayesian Computation with R", Jim Albert.
6. "Kickstarting R"- cran.r-project.org/doc/contrib/Lemon-kickstart.
7. "Introduction to R" -cran.r-project.org/doc/manuals/R-intro.html.

Course Code	Course Name	Category	L	P	T	Credit
	SQL AND NOSQL DATABASE MANAGEMENT SYSTEMS	C	4			4

Preamble

- To understand the basics of Database Management.
- To establish the need for SQL and NoSQL Databases in Data Analytics.
- To understand the different types of SQL and NoSql databases and their uses.

Prerequisite

- Basics of database , MS Access

Course Outcomes

On the successful completion of the course, students will be able to

Course Outcomes		Bloom's Level
CO1	Define the basic concepts of Databases	Remember, Understand
CO2	Identify the scope of Sql and NoSql Databases, advantages over conventional Databases.	Remember, Understand, Apply
CO3	Identify the right kind of SQL and NoSql Database for the right purpose while doing a Data Analytics project.	Apply, Analyze, Evaluate, Create
CO4	Work with a particular NoSql tool.	Apply, Evaluate, Create

Mapping with Programme Outcomes

COs	PSO1	PSO2	PSO3	PSO4	PSO5	PSO6	PSO7
CO1	S	S	S	M	M	L	L
CO2	S	S	S	M	S	L	L
CO3	S	S	S	M	S	L	L
CO4	M	S	S	S	S	L	S

Assessment Pattern

Bloom's Category	Continuous Internal Assessment (25)			Terminal Examination (75)
	I	II	III	
Remember	5	5	5	22
Understand	6	6	6	23
Apply	5	5	5	10
Analyze	5	5	5	10
Evaluate	2	2	2	5
Create	2	2	2	5

Syllabus

UNIT I Fundamental concepts: Introduction to DBMS- Database System Architecture – levels- Database users and DBA- Entity-Relationship model- E-R Diagram- Mapping- Translating E-R model into Relational model- Describe the term RDBMS- ACID- Distributed System. **(11hrs)**

UNIT II Relational Model: The relational Model, Relational Algebra- Fundamental operations- Additional Operations- SQL fundamentals- DDL- DML- DCL- Keys- Constraints- Nested Queries- Stored Procedures- Stored Functions. **(11hrs)**

UNIT III Introduction to NoSQL and MongoDB: Big Data- Characteristics- Types of Data- NoSQL Concept- Characteristics- CAP Theorem and BASE properties- Difference between NoSQL and RDBMS- Benefits of NoSQL- Categories of NoSQL- Architectural difference between the database types.
MongoDB: MongoDB Features- MongoDB vs RDBMS terminology/concepts- Advantages- Data Modelling- Create Database- Drop Database- Create Collection- Drop Collection.
(14hrs)

UNIT IV MongoDB Datatypes & Commands: Data types- Insert Document- Filter Document- Update Document- Projection- Limit Records- Sort Records- Indexing- Aggregation- Replication- Sharding- Create Backup- Deployment- MongoDB-Java (JDBC Driver).
(12hrs)

UNIT V Neo4j Graph Database: Introduction- Data model- Environment setup- Building blocks - Creating nodes and relationship- Write clause set- merge- delete- remove- for each - Read clause match- optional match- where- count. **(12hrs)**

Total: 60hrs

Reference Books and URLs:

1. “Database Management Systems Designing and Building Business Application”, G. V. Post, McGraw Hill International edition, 1999.
2. “Database Management Systems”, Raghu Ramakrishnan, WCB/McGraw Hill, 1998.
3. “An Introduction to Database Systems 7th Edition”, C.J. Date, Addison Wesley, 2000.
4. “eXist: A NoSQL Document Database and Application Platform”, Erik Siegel, Adam Retter, O'Reilly Media, 2014, ISBN: 978-1-44933-710-0
5. “Professional NoSQL”, Shashank Tiwari, Wrox-2011, ISBN: 047094224X,9780470942246
6. “The Definitive Guide to MongoDB: The NoSQL Database for Cloud and Desktop Computing”, Eelco Plugge, DUPTim Hawkins, Peter Membrey, Apress, 2010, ISBN: 1430230517, 9781430230519.
6. https://www.tutorialspoint.com/neo4j/neo4j_data_model.htm

Course Code	Course Name	Category	L	P	T	Credit
	DATA VISUALIZATION	C	4			4

Preamble

- To understand the need for Data Visualization.
- To understand the basic concepts of Data Visualization.
- To acquire skills in working with R for Visualization.

Prerequisite

- Basic scientific graphs, charts

Course Outcomes

On the successful completion of the course, students will be able to

Course Outcomes		Bloom's Level
CO1	Identify the need for Data Visualization in Data Analytics.	Remember, Understand
CO2	Identify a right tool for Data Visualization based on the Data Analytics Problem.	Apply, Analyze
CO3	Identify the right visualization technique to obtain information, knowledge and insight from particular Data-set.	Apply, Analyze, Evaluate
CO4	Execute Visualization tasks in R	Apply, Analyze, Evaluate, Create

Mapping with Programme Outcomes

COs	PSO1	PSO2	PSO3	PSO4	PSO5	PSO6	PSO7
CO1	S	S	M	M	L	L	L
CO2	S	S	S	M	S	L	L
CO3	M	M	M	M	S	L	L
CO4	M	M	M	S	S	L	L

Assessment Pattern

Bloom's Category	Continuous Internal Assessment (25)			Terminal Examination (75)
	I	II	III	
Remember	5	5	5	22
Understand	6	6	6	23
Apply	5	5	5	10
Analyze	5	5	5	10
Evaluate	2	2	2	5
Create	2	2	2	5

Syllabus

UNIT 1: Introduction to Data Visualization: Data- Information- Knowledge- Data Analysis and Insights- Transforming Data into Information- Transforming Information into Knowledge- Transforming Data into Insight- Data Visualization History- Data Visualization for decision making and Data Visualization Tools for Analytics. **(10hrs)**

UNIT II: Basic Data Manipulation in R: Introduction to R- Features- Installing and getting help in R- Data types- Variables - Operators- Decision making and looping statements- Functions -Data Manipulation: Strings, Vectors, Lists- Matrix- Arrays and Factors – Data Frame: Creating data frame, adding row and columns and statistical summary of data frame - Packages. **(14hrs)**

UNIT III Basic Data Visualization: A simple Line chart- Bar Plot- Pie Chart - Scatter Plot: Scatter Plot with texts, labels and lines, An interactive Scatter Plot - Box Plot – Plotting Multiple Curves- Lines function- Adding Legend and Text- histogram. **(14hrs)**

UNIT IV Graphic Visualization: Grammar of Graphics -The Setup for ggplot2 - Aesthetic Mapping in ggplot2 - Geometry in ggplot2 - Labels in ggplot2 - Themes in ggplot2 - ggplot2 Common Charts: Bar Chart, Density Plot, Scatterplot, Box plot - Interactive Charts with Plotly and ggplot2. **(12hrs)**

UNIT V Visualization of Statistical concepts: Mean – Median – Mode – Linear Regression – Multiple Linear Regression- Logistic Regression - Normal Distribution - Binomial Distribution- Poisson Distribution **(10hrs)**

Total: 60 hrs

Reference Books and URLs**:**

1. "Learn R for Applied Statistics With Data Visualizations, Regressions, and Statistics", Eric Goh Ming Hui, 2019.
2. "R Data Visualizing Codebook", Atmajithsingh Gohil, Packt PublishingLtd., UK.
3. "A Handbook of Statistical Analyses Using R", 2nd Edition, Brian S. Everitt and Torsten Hothorn, Tayor and Francis Group.
4. "The Art of R Programming", Normann Mattloff, No Starch Press, San Fransisco.
5. "Programming with Data", Chambers, J. M., USA: Springer, 1998.
6. "Visualizing Categorical Data", Meyer, D., Zeileis, A., Karatzoglou, A., and Hornik, K., 2006.
7. "A Little Book of R for Time Series", Avril Coglan, Wellcome Trust Sanger Institute, U.K.
8. http://CRAN.R-project.org, R package version 0.9-91.

Course Code	Course Name	**Category**	L	P	T	Credit
	SQL AND NOSQL DATABASE MANAGEMENT SYSTEMS LABORATORY	L		2		2

Exercises to improve your SQL, NoSQL Database Management skills

1. Create databases with constrains using SQL.
2. Design the database using DML Statements: Insert, Update, Delete and Select queries with different conditions using SQL.
3. Implement group by clause and its function using SQL.
4. Working with nested queries.
5. Create a procedure to get the emp details of person using procedure in SQL.
6. Create a function to update the salary of employee in SQL.
7. Implement database with suitable example using MongoDB.
8. Implement all basic operations: create, drop database and collection.
9. Implement the basic operations in collection: insert, delete, update.
10. Use MongoDB to process semi structured/unstructured data collections.
11. Aggregation and indexing with suitable example using MongoDB.
12. Indexing and querying with MongoDB using suitable example.
13. Connectivity with MongoDB using any Java application.
14. Create and delete node, relationship, index using Neo4j.
15. Implement write operations: set, merge, delete, remove, for each clause using Neo4j.
16. Implement read operations: match, optional match, where, count clause using Neo4j.

Course Code	Course Name	**Category**	L	P	T	Credit
	DATA VISUALIZATION LABORATORY	L		2		2

Exercises using R

1. Software and Package Installation Steps.
2. Loading different kinds of data formats.
3. Working with statistical functions.
4. Exercises to work with Vectors and function.
5. Matrix manipulation operations.
6. Working with conditional statement and loops.
7. Creating and summarizing data using data frame.
8. Working with CSV file.
9. Plotting Line and pie charts.
10. Scatter plot exercises with Text, Labels, Lines and Connecting Points.
11. Plotting categorical variables using different Bar Plots.
10. Working with Histograms plots.
12. Working with interactive plot using ggplot2.
13. Visualizing statistical concepts for Normal Distribution and Binomial Distribution.
14. Visualizing statistical concepts for Poisson Distribution.
15. Visualizing data using boxplot.

SEMESTER II

LIST OF COURSES

(For the Candidates Admitted From 2022-23 Onwards)

Sl. No.	Course code	Course name
1.		Supportive Course I
2.		Machine Learning
3.		Descriptive and Discovery Analytics
4.		Elective (Group A)
5.		Elective (Group A)
6.		Elective (Group A)
7.		Machine Learning Laboratory
8.		Descriptive and Discovery Analytics Laboratory

Course Code	Course Name	Category	L	P	T	Credit
	MACHINE LEARNING	C	4			4

Preamble

- To know the importance of Machine Learning
- To learn basic Machine learning algorithms.
- To understand how to apply the learning algorithms for various prediction problems.

Prerequisite

Introduction to Mathematical and Statistical concepts.

Course Outcomes

On the successful completion of the course, students will be able to

Course Outcomes		Bloom's Level
CO1	Know the importance of Machine Learning	Remember, Understand
CO2	Learn basic Machine learning algorithms	Remember, Understand
CO3	Understand how to apply the learning algorithms for various prediction problems	Apply, Analyze, Evaluate, Create.

Mapping with Programme Outcomes

COs	PSO1	PSO2	PSO3	PSO4	PSO5	PSO6	PSO7
CO1	S	S	M	M	L	L	L
CO2	S	S	M	M	L	L	L
CO3	M	S	S	S	L	L	M

Assessment Pattern

Category	Continuous Internal Assessment (25)			Terminal Examination (75)
	I	II	III	
Remember	5	5	5	22
Understand	6	6	6	23
Apply	5	5	5	10
Analyze	5	5	5	10
Evaluate	2	2	2	5
Create	2	2	2	5

Syllabus

UNIT I - Introduction : Learning Problems – Perspectives and Issues – Concept Learning – Version Spaces and Candidate Eliminations – Inductive bias – Decision Tree learning – Representation – Algorithm – Heuristic Space Search. **(12hrs)**

UNIT II Workflow and Types of Machine Learning Algorithms: Process of Machine Learning - Machine Learning Workflow– Types of Common Machine Learning Algorithms– Performance Metrics. **(12hrs)**

UNIT III Eessential Concepts for Machine Learning: Data Pre-processing– Feature Engineering– Regression Concepts– Classification algorithms– Clustering algorithms. **(14hrs)**

UNIT IV Instant Based Learning: K- Nearest Neighbour Learning – Locally weighted Regression – Self Organizing Map – Vector Quantization - Locally Weighted Learning. **(10hrs)**

UNIT V Advanced Learning: Neural Network Representation – Problems – Perceptrons – Multilayer Networks, Activation Functions, Gradient Descent Rule, Stochastic Gradient Descent Optimization, Back Propagation Algorithm – Genetic Algorithm -Basic concepts – Working Principles – Problems. **(12hrs)**

Total: 60 hrs

Reference Books and URLs**:**

1. “Machine Learning”, Tom M. Mitchell, McGraw-Hill Education (India) Private Limited, 2013.

2. “Introduction to Machine Learning (Adaptive Computation and Machine Learning)”, EthemAlpaydin, The MIT Press, 2004.

3. “Machine Learning: An Algorithmic Perspective, Stephen Marsland, CRC Press, 2009.

4. “Genetic Algorithms and Genetic Programming”, Michael Affenzeller, Stephan Winkler, Stefan Wagner, Andreas Beham, CRC Press Taylor and Francis Group.

Course Code	Course Name	Category	L	P	T	Credit
	DESCRIPTIVE AND DISCOVERY ANALYTICS	C	4			4

Preamble

- To understand the role of descriptive analytics on data.
- To acquire knowledge in different Descriptive and Discovery analytics concepts.
- To get the idea on which model to apply for descriptive and discovery analytics.

Prerequisite

Basics of Descriptive Statistics

Course Outcomes

On the successful completion of the course, students will be able to

Course Outcomes		Bloom's Level
CO1	Identify the scope, role and application of Descriptive Analytics.	Remember, Understand
CO2	Handle raw data using Python.	Apply, Analyze
CO3	Remember various descriptive and discovery analytics techniques.	Understand, Apply, Analyze, Evaluate
CO4	Identify the right kind of analytics model for the right purpose while doing a Data Analytics project.	Analyze, Evaluate, Create

Mapping with Programme Outcomes

COs	PSO1	PSO2	PSO3	PSO4	PSO5	PSO6	PSO7
CO1	S	S	M	M	L	L	L
CO2	M	M	M	S	S	L	L
CO3	S	S	M	M	M	L	L
CO4	M	M	M	S	M	L	L

Assessment Pattern

Category	Continuous Internal Assessment (25)			Terminal Examination (75)
	I	II	III	
Remember	5	5	5	22
Understand	6	6	6	23
Apply	5	5	5	10
Analyze	5	5	5	10
Evaluate	2	2	2	5
Create	2	2	2	5

Syllabus

Unit I - Introduction to Descriptive Analytics: Introduction to Descriptive Analytics – Examples - The Role of Descriptive Analytics in Future Data Analysis – An industry Application - Descriptive Data Collection: Survey Overview - Net Promoter Score and Self-Reports - Survey Design - Passive Data Collection - Media Planning- Causal Data Collection and Summary. **(14hrs)**

Unit II – Empowering data analysis with pandas: Data structure - Inserting and Exporting Data- Data Cleansing- Checking and Filling Missing Data- String Operations- Merging Data- Aggregation operations -Joins – Case Study. **(11hrs)**

Unit III – Basics of Discovery Analytics: Comparing two groups - Drawing inferences - Independent groups - Dependent groups - Categorical association - Chi-squared test for association - The Chi-squared test - Interpreting the Chi-squared test - Chi-squared test for goodness of fit - An alternative to the Chi-squared test- Case Study. **(13hrs)**

Unit IV - Simple and Multiple Regressions: Simple regression - Describing quantitative association - Drawing inferences - Pitfalls in regression - Testing the model - Checking assumptions - Exponential regression - Multiple regression – Model- Tests - Overall test - Individual tests - Checking assumptions. **(12hrs)**

Unit V - Parametric Tests and Non-parametric Tests: Basics and One-way ANOVA - Assumptions and F-test - Post-hoc t-tests - Factorial ANOVA - ANOVA and regression - Non-parametric tests - The basics - Comparing groups with respect to mean rank - Several samples - Kruskal-Wallis test. **(10hrs)**

Total: 60 hrs

Reference Books and URLs:

1. "Mastering Python for Data Science", Samir Madhavan, Packt, 2015.
2. https://www.coursera.org/learn/wharton-customer-analytics
3. http://www.dataversity.net/fundamentals-descriptive-analytics
4. https://www.coursera.org/learn/inferential-statistics

Course Code	Course Name	Category	L	P	T	Credit
	MACHINE LEARNING LABORATORY	L		2		2

Exercises using Python and R

1. Working with Python-Tutorials

2. Working with R- tutorials

3. Understanding Machine Learning Problems, Training Dataset, Test Data Set.

4. Execute Linear Regression in Python and R using suitable Training and Testing data set for predicting the cost of a flat.

5. Execute Logistic Regression in Python and R using suitable Training and Testing data to predict discrete outputs.

6. Execute Decision trees in Python and R using suitable Training and Testing data set for making suitable predictions.

7. Execute Support Vector Machine in Python and R using suitable Training and Testing data set for making suitable predictions.

8. Execute K means clustering in Python and R using suitable Training and Testing data set for making suitable predictions.

9. Execute Random Forest in Python and R using suitable Training and Testing data set for making suitable predictions.

10. Execute KNN in Python and R using suitable Training and Testing data set for making suitable predictions.

Reference and URLs:

1.https://www.youtube.com/watch?v=2uCXIbkbDSE

2. https://www.youtube.com/watch?v=eDrhZb2onWY

3.https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/

4.https://www.analyticsvidhya.com/blog/2016/04/complete-tutorial-tree-based-modeling-scratch-in-python/#one

5. https://www.analyticsvidhya.com/blog/2014/10/support-vector-machine-simplified/

Course Code	Course Name	Category	L	P	T	Credit
	DESCRIPTIVE AND DISCOVERY ANALYTICS LABORATORY	L		2		2

Exercises using Python

1. Program to import and export data files of different data format.
2. Program to apply your own or another library's functions to Pandas objects.
3. Program to identify and handle missing data values.
4. Program for data normalization.
5. Program to convert common standard data format.
6. Program to summarize the data.
7. Program to implement ANOVA.
8. Program to implement correlation.
9. Program to perform various SQL operations using pandas.
10. Program to Model Evaluation Using Visualization.
11. Program to Simple Linear Regression.
12. Program to Multiple Linear Regression model.
13. Program to R-squared and MSE for In-Sample Evaluation.
14. Program to perform Model Evaluation.

SEMESTER III

LIST OF COURSES

(For the Candidates Admitted From 2022-23 Onwards)

Sl. No.	Course Code	Course Name
1.		Supportive Course II
2.		Predictive and Prescriptive Analytics
3.		Ethics for Data Scientists
4.		Elective (Group B)
5.		Elective (Group B)
6.		Predictive and Prescriptive Analytics Laboratory
7.		Mini Project

Course Code	Course Name	Category	L	P	T	Credit
	PREDICTIVE AND PRESCRIPTIVE ANALYTICS	C	4			4

Preamble

- To get introduced to the need of Predictive and Prescriptive Analytics.
- To understand the concepts of text mining and time forecasting concepts for predictive analytics.
- To practice predictive techniques using Rapid Miner.

Prerequisite

Introduction to Machine Learning, Data Analytics

Course Outcomes

On the successful completion of the course, students will be able to

Course Outcomes		Bloom's Level
CO1	Understand the need of Predictive Analytics	Understand
CO2	Understand data mining and time forecasting concepts for making Predictions	Remember, Understand, Apply
CO3	Apply data mining and time forecasting techniques for predictions	Apply, Analyze, Evaluate, Create

Mapping with Programme Outcomes

Cos	PSO1	PSO2	PSO3	PSO4	PSO5	PSO6	PSO7
CO1	S	S	M	M	M	L	L
CO2	S	S	M	M	L	L	L
CO3	M	M	S	S	S	L	L

Assessment Pattern

Category	Continuous Internal Assessment (25)			Terminal Examination (75)
	I	II	III	
Remember	5	5	5	22
Understand	6	6	6	23
Apply	5	5	5	10
Analyze	5	5	5	10
Evaluate	2	2	2	5
Create	2	2	2	5

Syllabus

Unit I: Predictive Analytics: Predictive Analytics-Definition- Problems- Regression analysis for Demand Curve and for making Predictions- Making Predictions using a Data Set- Probability Models- Implementation of a Model. **(13hrs)**

Unit II: Data mining concepts for Predictions: Data Mining Process- Association analysis: Concepts of Mining Association Rules- Apriori Algorithm- Frequent Item Set Generation- FP Growth Algorithm. **(11hrs)**

Unit III: Text Mining: Text Mining - Working: Term Frequency- Inverse Document Frequency- Associated Terminologies- Text mining with Clustering and Classification- Case Studies- Key Word Clustering- Predicting the gender of Blog Authors. **(12hrs)**

Unit IV: Time Series Forecasting: Need- Data driven approaches: Naive Forecast, Simple Average, Moving Average, Weighted Moving Average, Exponential Smoothing - Model driven forecasting: Linear Regression, Polynomial Regression, Linear Regression with Seasonality. **(12hrs)**

Unit V: Prescriptive Analytics: Introduction- Prescriptive Analytics- Difference between Predictive and Prescriptive Analytics-Using the Data to Maximize Revenue- Parameters of the Model- Market Structure- Competition and Online Advertising Models. **(12hrs)**

Total: 60 hrs

Reference Books and URLs:

1. "Predictive Analytics and Data Mining Concepts and Practice with RapidMiner", 'Vijay Kotu, Bala Deshpande, PhD, Elsevier Publication.

2. https://www.coursera.org/learn/wharton-customer-analytics

3. https://www.analyticsvidhya.com/blog/2014/10/support-vector-machine-simplified/

4.https://www.neuraldesigner.com/blog/6_Applications_of_predictive_analytics_in_business_intelligence

5.https://doc.lagout.org/Others/Data%20Mining/Data%20Mining%20and%20Predictive%20Analytics%20%5BLarose%20%26%20Larose%202015-03-16%5D.pdf

Course Code	Course Name	Category	L	P	T	Credit
	ETHICS FOR DATA SCIENTISTS	C	4			4

Preamble

- To understand the benefits and drawbacks of using data while using them for making predictions.
- To understand the structure of Ethics, law and societal values.

Prerequisite

Introduction to Data Analytics

Course Outcomes

On the successful completion of the course, students will be able to

Course Outcomes		Bloom's Level
CO1	Understand the benefits and drawback of data.	Remember, Understand
CO2	Understand ethical and Societal values while using data.	Remember, Understand, Apply
CO3	Apply data ethics while practicing Data Science.	Understand, Apply, Analyze

Mapping with Programme Outcomes

COs	PSO1	PSO2	PSO3	PSO4	PSO5	PSO6	PSO7
CO1	M	M	L	S	S	S	L
CO2	M	M	L	S	S	S	S
CO3	S	S	L	S	S	S	S

Assessment Pattern

Category	Continuous Internal Assessment (25)			Terminal Examination (75)
	I	II	III	
Remember	5	5	5	25
Understand	10	10	10	25
Apply	5	5	5	15
Analyze	5	5	5	10
Evaluate				
Create				

Syllabus

Unit I: Introduction: Ethics- Definition -Need- Examples- Five Cs-Consent, Clarity, Consistency, Control, Institutional Review Board in India and US, Informed Consent- Importance, Limitations of informed Consent- Real World Cases and Discussion**. (14hrs)**

Unit: II Data Ownership and Privacy: Data Collection-Issues, Data ownership- Limits of Data Ownership, Limitations in using Data, Intellectual Property rights- Privacy Issues- Introduction, Degrees, Risks - Anonimity: De-Identification, redaction, pseudonymization, anonymization, Cases and Discussion **(12hrs)**

Unit: III Data Validity: Validity - Errors in data, Sources Errors in Data, Attributes for Research, Metrics for Research- Errors in Data Processing- Errors in Model Design, Issues on Society, Cases and Discussion. **(12hrs)**

Unit IV Fairness of Algorithms: Algorithm Fairness- Definition, Importance, Unfair Algorithms, Correct and Misleading results, Fairness in Machine Learning and Artificial Intelligence, Fairness enhancing mechanisms, Cases and Analysis. **(10hrs)**

Unit V Code of Ethics: Ethics during: Decision making, communicating with clients, handling data, Conflict of Interest. Staying Informed, Quality of Data, Quality of Evidence, Misconduct, Maintaining Integrity. **(12hrs)**

Total: 60 hrs

Reference Books and URLs:

1. “Ethics and Data Science”, DJ Patil, Hilary Mason, Mike Loukides, O'Reilly Media, Inc.

2. .“A Beginner’s Guide to Data Ethics”, Sophie Lou and Mark Yang, Published in Big Data at Berkeley.

3. MOOCS: https://courses.edx.org/courses/course v1:MichiganX+DS101x+1T2018/ courseware/94ac457869964552a69a3f37ba579954/671f4645836145eea658edbf9298be64/

4. https://en.wikipedia.org/wiki/Fairness_(machine_learning)

5.https://arxiv.org/pdf/2001.09784.pdf

6.https://towardsdatascience.com/a-gentle-introduction-to-the-discussion-on-algorithmic fairness-740bbb469b6

7.https://nvlpubs.nist.gov/nistpubs/ir/2015/nist.ir.8053.pdf

8.https://www.datascienceassn.org/sites/default/files/datasciencecodeofprofessionalconduct.pdf

Course Code	Course Name	Category	L	P	T	Credit
	PREDICTIVE AND PRESCRIPTIVE ANALYTICS LABORATORY	L		2		2

Exercises using Rapid Miner

1. Getting familiar with Rapid Miner.
2. The retrieval of association rules from a data set is implemented through the FP-Growth algorithm in RapidMiner.
3. Execute the **Crawl Web** operator of Rapid Miner to allow setting up of simple crawling rules and based on these rules to store the crawled pages in a directory for further processing.
4. Execute the **GetPage** operator of Rapid Miner to retrieve a single page and stores the content as an example set and the **Get Pages** operator to access multiple pages identified by their URLs contained in an input file.
5. Predict the content of Blog Authors using the contents of Blog using Rapid Miner.
6. Model a Time series to forecast profits using Rapid Miner for the next twelve months for a dataset which has the monthly profit of a particular product from January 2009 to June 2010. The process consists of the following three steps: (1) set up windowing; (2) train the model with several different algorithms; and (3) generate the forecasts.

Reference Books and URLs:

1. "Predictive Analytics and Data Mining Concepts and Practice with RapidMiner", Vijay Kotu, Bala Deshpande, PhD, Elsevier Publication. Exercise: 2, Chapter: 6 Pages 211-214, Exercises: 3,4 Chapter 9 Pages: 285, Exercises: 5 Chapter 9 Pages 287-302, Exercises; 6.Chapter 10-Pages: 318-326.

Course Code	Course Name	Category	L	P	T	Credit
	MINI PROJECT	P		3	1	4

The student should do a Mini Project in a topic from Data Analytics. The internal assessment is for 50 marks from the three mini project reviews. The external is for 50 marks from the Final Mini project report, presentation and viva-voce.

SEMESTER IV

LIST OF COURSES

(For The Candidates Admitted From 2022-23 Onwards)

Sl. No.	Subject code	Course name
1.		Elective 6 (E-Pathshala)
2.		Elective 7
3.		Elective 8
4.		Major Project

Course Code	Course Name	Category	L	P	T	Credit
	MAJOR PROJECT	P		7	3	10

The Student can carry out the final Major project work in 4th semester in an IT company, the student can get permission from the concerned Project Supervisor, course teachers and Head of the Department after submitting the Acceptance Letter from the IT Company. For Major Project, the CIA is carried out for 50% marks and the External Assessment is for 50% marks from the Final Project Presentation, Project Report, and Viva-Voce.

LIST OF ELECTIVE COURSES FOR SEMESTER II (GROUP A)

(For The Candidates Admitted From 2022-23 Onwards)

Sl. No.	Course code	Course name
1.		Programming with Python for Data Analytics
2.		Natural Language Processing
3.		Computer Vision and Applications
4.		Programming with Mangodb for Data Analytics
5.		Big Data Analytics on Genomic Data

Course Code	Course Name	Category	L	P	T	Credit
	PROGRAMMING WITH PYTHON FOR DATA ANALYTICS	E	3			3

Preamble

- To learn the fundamentals, functions and oops concepts of python.
- To implement various Visualization methods and Machine Learning models with python.

Prerequisite

Basic Programming Skills

Course Outcomes

On the successful completion of the course, students will be able to

Course Outcomes		Bloom's Level
CO1	Learn the functions and oops concepts of python.	Remember, Understand
CO2	Analyze various data sets using Visualization packages of python.	Apply, Analyze, Evaluate
CO3	Apply Machine Learning models with python	Apply, Analyze, Evaluate, Create

Mapping with Programme Outcomes

COs	PSO1	PSO2	PSO3	PSO4	PSO5	PSO6	PSO7
CO1	S	S	M	M	M	L	L
CO2	M	M	S	S	S	L	L
CO3	M	M	S	S	S	L	L

Assessment Pattern

Category	Continuous Internal Assessment (25)			Terminal Examination (75)
	I	II	III	
Remember	5	5	5	22
Understand	6	6	6	23
Apply	5	5	5	10
Analyze	5	5	5	10
Evaluate	2	2	2	5
Create	2	2	2	5

Syllabus

Unit: I Fundamentals of Python: Introduction to Python- Features of Python- Python Basics: Comments, Variables- Data Types: Numeric Data type - Lists: list operations, list slices, list methods, list loop, mutability, aliasing, cloning lists, list parameters- Tuples: tuple assignment, tuple as return value- Dictionaries: operations and methods- Set- Data frame- String- Type Conversion- Operators- Namespacing and Scopes. **(8hrs)**

Unit: II Control Flow, Functions and OOPS Concepts: Control and Looping Statement: if, elif else- Looping Statement: for, while, nested loops - Difference between object and procedural oriented programming- Classes and Objects - Object Oriented Programming methodologies: Abstraction, Inheritance, Polymorphism, Encapsulation. **(8hrs)**

Unit: III Numerical Programming: Numpy- Creating Numpy Array- Numpy Data objects- dtype -Numerical Operations on Numpy Arrays - Numpy Arrays: Concatenating, Flattening and Adding Dimensions - Python, Random Numbers and Probability - Weighted Probabilities - Synthetical Test Data With Python - Numpy: Boolean Indexing - Matrix Multiplication, Dot and Cross Product - Reading and Writing Data Files. **(8hrs)**

Unit: IV Visualization: Overview of Matplotlib-Various Plots: Line, Bar, Scatter plots, Pie chart and Histogram, Format Plots: Shading Regions with fill_between() - Spines and Ticks, Adding Legends and Annotations. Introduction to Ploly – Features – Package Structure - Exporting to Static Images – main packages: plotly.plotly, plotly.graph_objs and plotly.tools.
(9hrs)

Unit: V Machine Learning: Simple linear regression: Evaluating the fitness of a model with a cost function, Solving ordinary least squares for simple linear regression, Evaluating the model - Applying linear regression: Exploring the data, Fitting and evaluating the model, Extracting features from categorical variables using scikit - Binary classification with logistic regression: Spam filtering, Binary classification performance metrics, Accuracy, Precision and recall and calculating the F1 measure.
Nonlinear Classification and Regression with Decision Trees: Decision trees, Training decision trees- Selecting the questions- Information gain- Gini impurity- Decision trees with scikit-learn.
(12hrs)

Total: 45 hrs

Reference books and URLs:

1. "Think Python: How to Think Like a Computer Scientist'', 2nd edition, Updated for Python 3, Allen B. Downey, Shroff/O'Reilly Publishers, 2016.
2. "An Introduction to Python – Revised and updated for Python 3.2, Guido van Rossum and Fred L. Drake Jr, Network Theory Ltd., 2011.
3. "Mastering Machine Learning with scikit-learn", Gavin Hackeling , Published by Packt Publishing Ltd., Birmingham B3 2PB, UK, 2014.
4. https://www.edureka.co/blog/python-basics/
5. https://python-course.eu/numerical-programming/
6. https://python-course.eu/books/bernd_klein_python_data_analysis_a4.pdf

Course Code	Course Name	Category	L	P	T	Credit
	NATURAL LANGUAGE PROCESSING	E	3			3

Preamble

- To understand the basic concepts of Natural Language Processing.
- To extract information from Unstructured Text, Analyze linguistic structure in text.
- To implement the techniques for text based analytics using python.

Prerequisite

Basics of Python Programming

Course Outcomes

On the successful completion of the course, students will be able to

Course Outcomes		Bloom's Level
CO1	Understand the basic concepts of Natural Language Processing	Remember, Understand
CO2	Extract information from Unstructured Text, Analyze linguistic structure in text.	Understand, Apply, Analyze, Create
CO3	Apply the techniques for text based analytics	Apply, Analyze, Evaluate, Create

Mapping with Programme Outcomes

COs	PSO1	PSO2	PSO3	PSO4	PSO5	PSO6	PSO7
CO1	S	S	M	M	L	L	L
CO2	M	S	S	S	L	L	L
CO3	M	S	S	S	S	L	L

Assessment Pattern

Category	Continuous Internal Assessment (25)			Terminal Examination (75)
	I	II	III	
Remember	5	5	5	22
Understand	6	6	6	23
Apply	5	5	5	10
Analyze	5	5	5	10
Evaluate	2	2	2	5
Create	2	2	2	5

Syllabus

Unit I Introduction to Natural Language Processing: Introduction- Language Modeling: Grammar-based LM- Statistical LM - Regular Expressions- Finite-State Automata – English Morphology- Transducers for lexicon and rules- Tokenization- Detecting and Correcting Spelling Errors- Minimum Edit Distance. **(8hrs)**

Unit II Word Level Analysis: Unsmoothed N-grams- Evaluating N-grams- Smoothing- Interpolation and Backoff – Word Classes- Part-of-Speech Tagging- Rule-based- Stochastic and Transformation-based tagging- Issues in PoS tagging – Hidden Markov and Maximum Entropy models. **(8hrs)**

Unit III Building Feature based Grammar: Grammatical Dilemmas - Use of Syntax- Context-Free Grammar- Parsing with Context-Free Grammar- Dependencies and Dependency Grammar- Grammar Development- Grammatical Features- Processing Feature Structures- Extending a Feature-Based Grammar. **(9hrs)**

Unit IV Classifying Text: Supervised Classification- Essential example for Supervised Classification- Evaluation- Decision Trees- Naive Bayes Classifiers- Maximum Entropy Classifiers- Modeling Linguistic Patterns- Information Extraction- Chunking- Developing and Evaluating Chunkers - Recursion in Linguistic Structure - Named Entity Recognition- Relation Extraction. **(10hrs)**

Unit V Language Processing and Python: Programming Framework of Python: Texts as Lists of Words- Making Decisions and Taking Control- Reusing Code- Computing with Language: Texts and Words- Accessing Text Corpora Conditional Frequency Distributions- Lexical Resources- WordNet- Accessing Text from the Web and from Disk - Strings: Text Processing at the Lowest Level- Text Processing with Unicode - Regular Expressions for Detecting Word Patterns - Segmentation Formatting: From Lists to Strings **(10hrs)**

Total: 45 hrs

Reference books and URLs:

1. "Speech and Language Processing: An Introduction to Natural Language Processing-Computational Linguistics and Speech", Daniel Jurafsky- James H. Martin, Pearson Publication, 2014.

2. "Natural Language Processing with Python", First Edition, Steven Bird, Ewan Klein and Edward Loper, O'Reilly Media, 2009.

3. "Handbook of Natural Language Processing: Second Edition", Nitin Indurkhya and Fred J. Damerau, Chapman and Hall/CRC Press, 2010.

4. "Natural Language Processing and Information Retrievall", Tanveer Siddiqui, U.S. Tiwary, Oxford University Press, 2008.

5. "Natural Language Processing with Python", Steven Bird, Ewan Klein, Edward Loper, O'Reilly Media Inc.

Course Code	Course Name	Category	L	P	T	Credit
	COMPUTER VISION AND APPLICATIONS	E	3			3

Preamble

- To understand important digital image processing operations for Computer Vision System.
- To understand different stages in the design of the computer vision system.
- To design Computer Vision based systems for various recognition applications.

Prerequisite

Basics of Mathematics and Digital Image Processing

Course Outcomes

On the successful completion of the course, students will be able to

Course Outcomes		Bloom's Level
CO1	Understand important digital image processing operations for Computer Vision System.	Remember, Understand
CO2	Understand different stages in the design of the computer vision system.	Understand, Apply, Analyze
CO3	Design Computer Vision based systems for various recognition applications.	Apply, Analyze, Evaluate, Create

Mapping with Programme Outcomes

COs	PSO1	PSO2	PSO3	PSO4	PSO5	PSO6	PSO7
CO1	S	S	M	M	L	L	L
CO2	S	S	M	M	L	L	L
CO3	M	M	S	S	L	L	L

Assessment Pattern

Category	Continuous Internal Assessment (25)			Terminal Examination (75)
	I	II	III	
Remember	5	5	5	22
Understand	6	6	6	23
Apply	5	5	5	10
Analyze	5	5	5	10
Evaluate	2	2	2	5
Create	2	2	2	5

Syllabus

Unit I Introduction: Computer Vision-brief history– Digital Image Fundamentals– Stages of Computer Vision– Acquisition– Pre-processing–Image Formation– Medium level Processing– High level Processing. **(10hrs)**

Unit II Basic Image Processing Operations: Filtering Operations– Morphological operations– Binary shape Analysis– Chain codes– Geometric primitives and Transformations– Exercises. **(10hrs)**

Unit III Model Fitting and Optimization: Model fitting and optimization– Scattered data interpolation–Variational methods and regularization– Markov random fields. **(10hrs)**

Unit IV Feature Detection, Matching and Recognition: Points and patches – Edges and contours– Contour tracking– Segmentation- Points and patches –Lines and vanishing points– Instance recognition – Image classification – Object detection – Semantic segmentation – Video understanding –Vision and language. **(10hrs)**

Unit V Computer Vision Applications: Vehicle Classification– Parking Slot Detection– Irrigation Management– Face Detection– Detect violent and dangerous situations– Automated mask detection. **(5 hrs)**

Total: 45 hrs

Reference Books and URLs:

1. "Digital Image Processing", Rafael Gonzalez, Richard Woods, 4th edition, Pearson, 2017.

2. "Computer Vision: Algorithms and Applications", Richard Szeliski, 2nd Edition, Springer, 2022.

3. "Computer vision and Artificial Intelligence techniques Applied to Robot Soccer", Alexander Baratella, Mauricio Gomes, International Journal of Innovative Computing, Information and Control.

4. "Machine Learning in Computer Vision", N. Sebe , Ashutosh Garg and Thomas S. Huang, Published by Springer.

5. "Computer Vision: Algorithms and Applications", Richard Szeliski, Published by Springer.

6. https://pdfs.semanticscholar.org/05de/dee0bcf1f0e73876537c7b1ee27ad769f695.pdf

7. https://infolab.usc.edu/csci599/Fall2002/paper/DML2_streams-issues.pdf

8. https://www.iotforall.com/computer-vision-applications-in-daily-life/

Course Code	Course Name	Category	L	P	T	Credit
	PROGRAMMING WITH MONGODB FOR DATA ANALYTICS	E	3			3

Preamble

- To understand the NoSQL databases, design goals, operations performing with java driver.
- To learn GSON and java EE environment.
- To create various applications.

Prerequisite

SQL and NoSQL Database Management Systems

Course Outcomes

On the successful completion of the course, students will be able to

Course Outcomes		Bloom's Level
CO1	Understand the NoSQL databases, design goals, operations performing with java driver	Remember, Understand
CO2	Learn GSON and java EE environment	Remember, Understand, Apply
CO3	Create various applications	Apply, Analyze, Evaluate, Create

Mapping with Programme Outcomes

COs	PSO1	PSO2	PSO3	PSO4	PSO5	PSO6	PSO7
CO1	S	S	M	M	M	L	L
CO2	S	S	M	M	S	L	L
CO3	M	M	S	S	S	L	L

Assessment Pattern

Category	Continuous Internal Assessment (25)			Terminal Examination (75)
	I	II	III	
Remember	5	5	5	22
Understand	6	6	6	23
Apply	5	5	5	10
Analyze	5	5	5	10
Evaluate	2	2	2	5
Create	2	2	2	5

Syllabus

UNIT I Introduction To MongoDB: NoSQL - MongoDB core elements- Installing and starting MongoDB- MongoDB tools - Introduction to the MongoDB shell- Securing database access. **(8hrs)**

UNIT II Java Driver for MongoDB: Mongo JDBC driver-creating, inserting, querying, updating, deleting - performing operation-listing collections- Java driver version 3-managing collections-inserting data-querying-updating-deleting documents. **(8hrs)**

UNIT III MongoDB Crud: MongoDB through java lens-extending mongoDB core classes-Gson API with MongoDb- Indexes in an applications-defining an index in java classes-compound indexes-text indexes- coding bulk operations-comparing plain inserts with BulkWrite Operations. **(9hrs)**

UNIT IV MongoDB in the JAVA EE 7 Enterprise Environment: Java EE land-java EE container-downloading wildfly-starting wildfly and testing the installation, designing application-designing the schema, building enterprise project with NetBeans-configuring wildfly-creating project-adding java classes-compiling and deploying the project-running, exposing the application. **(10hrs)**

UNIT V Managing Data Persistence with MongoDB and JPA: An overview of the java persistence API-entering hibernate OGM-building a project using hibernate OGM-using native queries in hibernate OGM- spring boot- spring data-mongo template component to access MongoDB **(10hrs)**

Total (45hrs)

Reference books and URLs:

1. "MongoDb for java developers", Francesco Marchioni, PACKT publishing, 2015.
2. "Java persistance with hibernate", Christian Bauer,Gavin King, 2007.
3. http://www.packtpub. com/jboss-as-5-development/book
4. https://nptel.ac.in/

Course Code	Course Name	Category	L	P	T	Credit
	BIG DATA ANALYTICS ON GENOMIC DATA	E	3			3

Preamble

- To articulate the main concepts of gene molecular structure.
- To learn algorithms to do gene sequencing and other relevant operations.
- To design suitable algorithms for gene based analytics.

Prerequisite

Basic knowledge of Genetics

Course Outcomes

On the successful completion of the course, students will be able to

Course Outcomes		Bloom's Level
CO1	Understand the main concepts of gene molecular structure.	Remember, Understand
CO2	Learn algorithms to do gene sequencing and other relevant operations.	Remember, Understand
CO3	Design suitable algorithms for gene based analytics.	Apply, Analyze, Evaluate, Create

Mapping with Programme Outcomes

COs	PSO1	PSO2	PSO3	PSO4	PSO5	PSO6	PSO7
CO1	S	S	M	M	L	L	L
CO2	S	S	M	M	L	L	L
CO3	M	S	S	S	L	L	L

Assessment Pattern

Category	Continuous Internal Assessment (25)			Terminal Examination (75)
	I	II	III	
Remember	5	5	5	22
Understand	6	6	6	23
Apply	5	5	5	10
Analyze	5	5	5	10
Evaluate	2	2	2	5
Create	2	2	2	5

Syllabus

Unit: I Molecular Biology Primer: Genetic Material- Genes- Molecule Codes for Genes- Structure of DNA- Information between DNA and Proteins- Components of Proteins- Analyze DNA, Copying DNA- Cutting and Pasting DNA- Measuring DNA Length- Probing DNA- Individuals of a Species Differ- Differences between Species- Bioinformatics.

Exhaustive Search: Restriction Mapping- Impractical Restriction Mapping Algorithms- A Practical Restriction Mapping Algorithm- Regulatory Motifs in DNA Sequences- Profiles- The Motif Finding Problem- Search Trees- Finding Motifs- Finding a Median String (**10hrs)**

Unit: II Greedy Algorithms: Genome Rearrangements- Sorting by Reversals- Approximation Algorithms- Breakpoints: A Different Face of Greed, A Greedy Approach to Motif Finding.
Dynamic Programming Algorithms: The Power of DNA Sequence Comparison- The Change Problem Revisited- The Manhattan Tourist Problem- Edit Distance and Alignments- Longest Common Subsequences- Global Sequence Alignment- Scoring Alignments- Local Sequence Alignment- Alignment with Gap Penalties- Multiple Alignment- Gene Prediction- Statistical Approaches to Gene Prediction- Similarity-Based Approaches to Gene Prediction- Spliced Alignment. **(10hrs)**

Unit: III Divide-and-Conquer Algorithms: Divide and Conquer Approach to Sorting- Space Efficient Sequence Alignment- Block Alignment and the Four Russians Speedup- Constructing Alignments in Subquadratic Time. **(8hrs)**

Unit: IV Graph Algorithms: Graphs- Graphs and Genetics- DNA Sequencing- Shortest Superstring Problem- DNA Arrays as an Alternative Sequencing Technique- Sequencing by Hybridization- SBH as a Hamiltonian Path Problem - SBH as an Eulerian Path Problem- Fragment Assembly in DNA Sequencing- Protein Sequencing and Identification- The Peptide Sequencing Problem- Spectrum Graphs - Protein Identification via Database Search- Spectral Convolution- Spectral Alignment. **(9hrs)**

Unit: V Pattern Matching and Clustering: Repeat Finding- Hash Tables- Exact Pattern Matching- Keyword Trees- Suffix Trees- Heuristic Similarity Search Algorithms- Clustering and Trees- Gene Expression Analysis - Hierarchical Clustering- k-Means Clustering- Clustering and Corrupted Cliques. **(8hrs)**

Total: 45 hrs

Reference books and URLs:

1. “Bioinformatics algorithms”, Neil c. Jones and pavel a. Pevzner, A Bradford Book, The MIT Press.
2. http://www.bioalgorithms.info

LIST OF ELECTIVE COURSES FOR SEMESTER III (GROUP B)

(For the Candidates Admitted From 2022-23 Onwards)

Sl. No.	Course code	Course name
1.		Big Data Security
2.		Deep Learning
3.		Data Mining Essentials for Data Analytics
4.		Data Visualization using Tableau
5.		Programming with Cassandra

Course Code	Course Name	Category	L	P	T	Credit
	BIG DATA SECURITY	E	3			3

Preamble

- To articulate the main concepts of big data privacy and security.
- To learn about hadoop ecosystem security and data security.

Prerequisite

Introduction to Big Data, Networking

Course Outcomes

On the successful completion of the course, students will be able to

Course Outcomes		Bloom's Level
CO1	Understand the main concepts of big data privacy and security	Remember, Understand
CO2	Learn about the hadoop ecosystem security.	Remember, Understand
CO3	Generate new ideas and innovations in data security	Apply, Evaluate, Create

Mapping with Programme Outcomes

COs	PSO1	PSO2	PSO3	PSO4	PSO5	PSO6	PSO7
CO1	S	S	M	M	L	M	L
CO2	S	S	M	L	M	M	L
CO3	M	S	M	S	S	M	L

Assessment Pattern

Category	Continuous Internal Assessment (25)			Terminal Examination (75)
	I	II	III	
Remember	5	5	5	25
Understand	10	10	10	25
Apply	5	5	5	10
Analyze				
Evaluate	2	2	2	5
Create	3	3	3	10

Syllabus

UNIT I – Big Data Privacy, Ethics And Security: Privacy – Reidentification of Anonymous People –Big Data Privacy self-regulating – Ethics – Ownership – Ethical Guidelines – Big Data Security – Organizational Security. **(8hrs)**

UNIT II - Security, Compliance, Auditing, and Protection: Steps to secure big data – Classifying Data – Protecting – Big Data Compliance – Intellectual Property Challenge – Research Questions in Cloud Security – Open Problems. **(10hrs)**

UNIT III – **Hadoop Security Design:** Kerberos – Default Hadoop Model without security - Hadoop Kerberos Security Implementation & Configuration. **(8hrs)**

UNIT IV – **Hadoop Ecosystem Security:** Configuring Kerberos for Hadoop ecosystem components – Pig, Hive, Oozie, Flume, HBase, Sqoop. **(9hrs)**

UNIT V – **Data Security & Event Logging:** Integrating Hadoop with Enterprise Security Systems - Securing Sensitive Data in Hadoop – SIEM system – Setting up audit logging in hadoop cluster. **(10hrs)**

Total: 45 hrs

Reference books and URLs:

1. "Think Bigger: Developing a Successful Big Data Strategy for Your Business", Mark Van Rijmenam, Amazon, 1 edition, 2014.

2. "Big Data Analytics: Turning Big Data into Big Money", Frank Ohlhorst John Wiley & Sons, John Wiley & Sons, 2013.

3. "Large Scale and Big Data: Processing and Management", Sherif Sakr, CRC Press, 2014.

4. "Securing Hadoop", Sudeesh Narayanan, Packt Publishing, 2013.

5. "Hadoop Security Protecting Your Big Data Problem", Ben Spivey, Joey Echeverria, O'Reilly Media, 2015.

Course Code	Course Name	Category	L	P	T	Credit
	DEEP LEARNING	E	3			3

Preamble

- To articulate the main concept of Deep Learning.
- To learn about CNN, RNN
- To build Deep Learning Applications.

Prerequisite

Programming, Statistics, Linear Algebra, Probability, Neural Networks.

Course Outcomes

On the successful completion of the course, students will be able to

Course Outcomes		Bloom's Level
CO1	Understand the main concepts of Deep Learning	Remember, Understand
CO2	Learn about CNN and RNN	Remember, Understand
CO3	Build Deep Learning Applications	Apply, Analyze, Evaluate, Create

Mapping with Programme Outcomes

COs	PSO1	PSO2	PSO3	PSO4	PSO5	PSO6	PSO7
CO1	S	S	M	L	L	L	L
CO2	S	S	M	M	L	L	L
CO3	M	S	S	S	M	L	L

Assessment Pattern

Category	Continuous Internal Assessment (25)			Terminal Examination (75)
	I	II	III	
Remember	5	5	5	22
Understand	6	6	6	23
Apply	5	5	5	10
Analyze	5	5	5	10
Evaluate	2	2	2	5
Create	2	2	2	5

Syllabus

UNIT I – Introduction: Multilayer Perceptron – Back Propagation Network–Stocatic Gradient Descent-Depth of Neural Networks-Neural Networks and Deep Learning – Difference-Activation Fuctions-RELU, LRELU, ERELU. **(10hrs)**

UNIT II – Autoencoders: Autoencoders-Definition- Characteristics-Parameters of Autoencoders- Relation with PCA-Architecture of Autoencoders–Training Auto encoders-Regularization in Autoencoders-Types-Contractive Autoencoder-Sparse Autoencoder-Denoising Autoencoder-Variational Autoencoder-Applications. **(10hrs)**

UNIT III – Convolutional Neural Networks [CNN]: CNN– General Architecture-layers-Filters-Working of CNN, Classic Networks- LeNet-5, AlexNet, ResNet, PixelNet-Applications. **(9hrs)**

UNIT IV – Recurrent Neural Networks: Bidirectional RNN, Encoder-Decoder Sequence to Sequence Architecture- Attention Mechanism-Back Propagation through Time for training RNN- Long Short Term Memory Networks. **(8hrs)**

UNIT V – Deep Learning Applications using PYTORCH: Building blocks of Neural Networks - Data Pre-processing-Deep Learning for Computer Vision-Classifying Dogs and Cats-Building Sentiment Classifier-Creating ResNet Model-Creating custom Pytorch dataset. **(8hrs)**

Total: 45 hrs

Reference Books and URLs:

1. "Deep Learning with Pytorch", Vishnu Subramaniam, Packt Publishing Ltd., UK, 2018. ISBN:978-1-78862-433-6.

2. https://nptel.ac.in/courses/106106184

3. https://www.edureka.co/blog/autoencoders-tutorial/

4. http://ufldl.stanford.edu/tutorial/unsupervised/Autoencoders/

5. https://www.v7labs.com/blog/autoencoders-guide

6. https://www.freecodecamp.org/news/convolutional-neural-network-tutorial-for-beginners/

7. https://data-flair.training/blogs/convolutional-neural-networks-tutorial/

8. https://training.galaxyproject.org/training-material/topics/statistics/tutorials/CNN/tutorial.html

9. https://www.youtube.com/watch?v=c36lUUr864M

Course Code	Course Name	Category	L	P	T	Credit
	DATA MINING ESSENTIALS FOR DATA ANALYTICS	E	3			3

Preamble

- To learn the concepts of data base technology evolutionary path which has lead to the need for data mining and its application.
- To Examine the types of the data to be mined and present a general classification of tasks and primitives to integrate a data mining system.

Prerequisite

Introduction to DBMS

Course Outcomes

On the successful completion of the course, students will be able to

Course Outcomes		Bloom's Level
CO1	Remember the concepts of data base technology evolutionary path which has lead to the need for data mining and its application.	Remember, Understand
CO2	Examine the types of the data to be mined and present a general classification of tasks and primitives to integrate a data mining system.	Understand, Apply, Analyze, Evaluate
CO3	Apply preprocessing statistical methods for any given raw data.	Apply, Analyze, Evaluate, Create

Mapping with Programme Outcomes

COs	PSO1	PSO2	PSO3	PSO4	PSO5	PSO6	PSO7
CO1	S	S	M	M	L	L	L
CO2	M	M	S	S	L	L	L
CO3	M	M	S	S	L	L	L

Assessment Pattern

Category	Continuous Internal Assessment (25)			Terminal Examination (75)
	I	II	III	
Remember	5	5	5	22
Understand	6	6	6	23
Apply	5	5	5	10
Analyze	5	5	5	10
Evaluate	2	2	2	5
Create	2	2	2	5

Syllabus

UNIT-I-Data Mining and Analysis: Data Matrix-Attributes-Data: Algebraic and Geometric View-Data: Probabilistic View-Data Mining. **(10hrs)**

UNIT-II Data Analysis Foundations: Numeric Attributes-Univariate Analysis-Bivariate Analysis-Multivariate Analysis-Data Normalization-Normal Distribution-Categorical Attributes-Graph Data-High-Dimensional Data-Dimensionality Reduction. **(10hrs)**

UNIT-III Frequent Pattern Mining: Itemset Mining-Summarizing Itemsets-Sequence Mining-Graph Pattern Mining. **(8hrs)**

UNIT-IV Clustering: Representative based Clustering-Hierarchical Clustering-Density based Clustering-Spectral and Graph Clustering-Clustering Validation. **(9hrs)**

UNIT-V Classification: Probabilistic Classification-Decision Tree Classifier-Linear Discriminant Analysis-Support Vector Machines. **(8hrs)**

Total: 45hrs

Reference books and URLs:

1. "Data Mining and Analysis: Fundamental Concepts and Algorithms", Mohammed J. Zaki, Wagner Meira, Jr., Cambridge University Press, May 2014. ISBN: 9780521766333.

2. https://repo.palkeo.com/algo/information-retrieval/Data%20mining%20and%20analysis.pdf/

Course Code	Course Name	Category	L	P	T	Credit
	DATA VISUALIZATION USING TABLEAU	E	3			3

Preamble

- To understand the importance of data visualization.
- To develop key data visualization skill set using Tableau.

Prerequisite

Data Visualization

Course Outcomes

On the successful completion of the course, students will be able to

Course Outcomes		Bloom's Level
CO1	Know the importance of data visualization	Remember, Understand
CO2	Develop key data visualization skill set	Apply, Analyze, Evaluate, Create
CO3	Apply and use Tableau for data visualization	Apply, Analyze, Evaluate, Create

Mapping with Programme Outcomes

COs	PSO1	PSO2	PSO3	PSO4	PSO5	PSO6	PSO7
CO1	S	S	M	M	M	L	L
CO2	M	M	S	S	S	L	L
CO3	M	M	S	S	S	L	L

Assessment Pattern

Category	Continuous Internal Assessment (25)			Terminal Examination (75)
	I	II	III	
Remember	5	5	5	22
Understand	6	6	6	23
Apply	5	5	5	10
Analyze	5	5	5	10
Evaluate	2	2	2	5
Create	2	2	2	5

Syllabus

Unit: I: Introduction: Data Visualization & its importance – Introduction of Tableau – Installation – Data types – Files types. **(10hrs)**

Unit: II: Design Flow: Types of visualization – Menu -File, Data, Worksheet, Dashboard, Story, Analysis, Map, Format, Server, Window – Help – Data source – Data view – Extracting data analytics. **(10hrs)**

Unit: III Operations: Operation – Editing – Meta data – Data joining- Data blending- Worksheet -Add, Remove, Update – Calculation - Operators, Functions, Numeric, String, Data, Table, LOD expression. **(10hrs)**

Unit: IV Visualization: Sets and Group- Filter -Sorting, Basic, Quick, Context, Condition, Top, Filter operation – Chart -pie, bar, scatter, tree map, box, bubble, gannt - Analysis. **(10hrs)**

Unit: V Dashboard: Trend line – formatting- forecasting- Dashboard- create, update, remove – story - creating and updating the story - removing the story. **(5hrs)**

Total: 45 hrs

Reference Books and URLs:

1. https://www.tableau.com/learn/training/20214
2. https://www.tutorialspoint.com/tableau/index.htm

Course Code	Course Name	Category	L	P	T	Credit
	PROGRAMMING WITH CASSANDRA	E	3			3

Preamble

- To understand the difference between RDBMS and Cassandra.
- To do various DBMS operations using Cassandra.

Prerequisite

Introduction to DBMS

Course Outcomes

On the successful completion of the course, students will be able to

Course Outcomes		Bloom's Level
CO1	Remember the difference between RDBMS and Cassandra	Remember, Understand
CO2	Do various DBMS operations using Cassandra	Understand, Apply, Analyze, Create
CO3	Link Java and Cassandra	Apply

Mapping with Programme Outcomes

COs	PSO1	PSO2	PSO3	PSO4	PSO5	PSO6	PSO7
CO1	S	S	M	M	M	L	L
CO2	M	M	S	S	S	L	L
CO3	M	M	M	M	M	L	L

Assessment Pattern

Category	Continuous Internal Assessment (25)			Terminal Examination (75)
	I	II	III	
Remember	5	5	5	22
Understand	8	8	8	23
Apply	5	5	5	10
Analyze	5	5	5	10
Evaluate				
Create	2	2	2	10

Syllabus

UNIT-I-CASSANDRA Basics: Introduction - NoSQL Database - NoSQL vs. Relational Database - Apache Cassandra - Features of Cassandra - History of Cassandra – Architecture - Data Replication in Cassandra - Components of Cassandra - Cassandra Query Language - Data Model – Cluster - Data Models of Cassandra and RDBMS-Installation -. Referenced API.

(11hrs)

UNIT-II CASSANDRA CQLSH: Cqlsh Commands-Shell Commands-KEYSPACE OPERATIONS-Create KeySpace -Alter KeySpace-Drop KeySpace. **(9hrs)**

UNIT-III Table Operations: Create Table - Alter Table - Drop Table - Truncate Table - Create Index - Drop Index - Batch Statements. **(8hrs)**

UNIT-IV CURD Operations: Create Data-Creating Data in a Table-Creating Data using Java API-Update Data-Read data-Reading Data using Select Clause-Where Clause-Reading Data using Java API - Delete data. **(9hrs)**

UNIT-V CQL Types: CQL data types-collection types-CQL collections-LIST,SET, MAP-CQL user defined datatypes-Creating,altering and deleting user defined data types. **(8hrs)**

Total: 45 hrs

Reference books and URLs:

1**. "**Mastering Apache Cassandra 3.x", Aaron Ploetz, Tejaswi Malepati, Nishant Neeraj.

2. https://www.gocit.vn/files/Cassandra.The.Definitive.Guide-www.gocit.vn.pdf

3. https://www.guru99.com/cassandra-tutorial.html

4. https://www.javatpoint.com/cassandra-setup-and-installation

LIST OF ELECTIVE COURSES FOR SEMESTER IV (GROUP C)

(For The Candidates Admitted From 2022-23 Onwards)

Sl. No.	Course code	Course name
1.		Cloud Computing [E-PG Pathshala]
2.		HADOOP for Data Analytics
3.		Cloud Platforms in Industry
4.		Storm for Data Analytics
5.		Spark for Data Analytics

Course Code	Course Name	Category	L	P	T	Credit
	CLOUD COMPUTING [E-PG PATHSHALA]	E	3			3

Preamble

- To articulate the main concepts, key technologies, strengths, and limitations of cloud computing and the possible applications for state-of-the-art cloud computing.
- To learn the core issues of cloud computing such as security, privacy, and interoperability.

Prerequisite

Introduction to networking

Course Outcomes

On the successful completion of the course, students will be able to

Course Outcomes		Bloom's Level
CO1	Remember the main concepts, key technologies, strengths, and limitations of cloud computing	Remember, Understand
CO2	Know the core issues of cloud computing such as security, privacy, and interoperability	Remember, Understand, Apply
CO3	Create new ideas and innovations in cloud computing	Apply, Analyze, Evaluate, Create

Mapping with Programme Outcomes

COs	PSO1	PSO2	PSO3	PSO4	PSO5	PSO6	PSO7
CO1	S	S	M	M	L	L	L
CO2	S	S	M	M	L	L	L
CO3	M	M	S	S	L	L	L

Assessment Pattern

Category	Continuous Internal Assessment (25)			Terminal Examination (75)
	I	II	III	
Remember	5	5	5	22
Understand	6	6	6	23
Apply	5	5	5	10
Analyze	5	5	5	10
Evaluate	2	2	2	5
Create	2	2	2	5

Syllabus

UNIT I Introduction to Cloud Computing: Evolution of Computing Paradigms-Utility Computing-Cloud Characteristics-Cloud Architecture-Delivery Models-Software-as-a-Service or SaaS, Platform-as-a-Service or PaaS , Infrastructure-as-a-Service or IaaS-Cloud Deployment Models-cloud storage-cloud security. **(10hrs)**

UNIT II Introduction to Distributed Systems: Distributed Systems and Distributed Computing-Benefits of Distributed Systems-Design Challenges of Distributed Systems-Distributed Systems and Cloud - Communication in Distributed Systems-Distributed Communication Paradigms-Message Passing- Remote Procedure Call-XML-RPC.**(10hrs)**

UNIT III Time Ordering & Replication: Need for Synchronization - The Problem- The Global Clock- The Global Clock Problem – Example- Synchronization Algorithms- Berkeley's Algorithms,Logical Timestamp of Events -Cloud Election- Bully Algorithm, Ring algorithm-Replication in Distributed Systems- Replication in Cloud- Replication Consistency -Replication System Architecture-Replication Types. **(10hrs)**

UNIT IV Virtualization: Problem with Traditional Systems- A Data Center- Virtualization Technology- Definition , Virtualization Terminologies , Goals of Virtualization,Types of Virtualization- Virtual Machines (VMs)- Virtual Machine Monitor -Virtualization and Emulation-Methods of Virtualization. **(8 hrs)**

UNIT V Web Services: Service Oriented Architecture- Services - Web Services Characteristics-Goals of Web Services - Types of Web Services - Components of Web Services Implementation-SOA and Cloud Computing- Resource Oriented Architecture (ROA)- RESTful Web Service. **(7hrs)**

Total: 45 hrs

Reference books and URLs:

1. https://epgp.inflibnet.ac.in/ahl.php?csrno=7

Course Code	Course Name	Category	L	P	T	Credit
	HADOOP FOR DATA ANALYTICS	E	3			3

Preamble

- To get introduced to the basic of Apache Hadoop.
- To learn about Hadoop's architecture and it's core components, such as Map Reduce and the Hadoop Distributed File System (HDFS).

Prerequisite

Introduction to Computer Architecture

Course Outcomes

On the successful completion of the course, students will be able to

Course Outcomes		Bloom's Level
CO1	Remember Hadoop's architecture and core components	Remember, Understand
CO2	Know about Hadoop clusters	Remember, Understand
CO3	Design big data solutions using Hadoop ecosystem	Apply, Analyze, Evaluate, Create

Mapping with Programme Outcomes

COs	PSO1	PSO2	PSO3	PSO4	PSO5	PSO6	PSO7
CO1	S	S	M	M	L	L	L
CO2	S	S	M	M	L	L	L
CO3	M	M	S	S	M	L	L

Assessment Pattern

Category	Continuous Internal Assessment (25)			Terminal Examination (75)
	I	II	III	
Remember	5	5	5	22
Understand	6	6	6	23
Apply	5	5	5	10
Analyze	5	5	5	10
Evaluate	2	2	2	5
Create	2	2	2	5

Syllabus

Unit I – Introduction: Hadoop - Definition- Big Data- Definition, Types- Sources-Open source software related to Hadoop- Big Data solutions working on Cloud- Deploying Hadoop. **(10hrs)**

Unit II – Hadoop Architecture: Hadoop components- Working of HDFS - List data access –patterns for which HDFS is designed- HDFS clusters. **(10hrs)**

Unit III– Hadoop Administration: Add and remove nodes from a cluster-Verify the health of a cluster- Start and stop clusters components - Modify Hadoop configuration parameters- Setup a rack topology. **(9hrs)**

Unit IV - Hadoop Map Reduce Framework: Overview of Map Reduce Framework- Map Reduce Architecture – Job tracker and Task tracker- Use cases of Map Reduce, - Anatomy of Map Reduce Program.- moving data using Flume and sqoop into Hadoop- scheduling using Oozie- job execution control in Hadoop. **(8hrs)**

Unit V – MapReduce Programs: Basic Map Reduce API Concepts- Writing Map Reduce Driver- Mappers and Reducers in Java-Speeding up Hadoop Development by Using Eclipse- Unit Testing Map Reduce Programs. **(8hrs)**

Total: 45 hrs

Reference Books and URLs:

1. **"**Hadoop Beginner's Guide**",** Garry Turkington, Published by Packt Publishing Ltd. Livery Place35 Livery Street Birmingham B3 2PB, UK.ISBN 978-1-84951-7-300
2. **"**Hadoop Illuminated", Mark Kerzner and Sujee Maniyam. [https://github.com/hadoop-illuminated/hadoop-book]
3. https://www.youtube.com/watch?v=-65WgvIJ5xo&feature=youtu.be
4. https://www.youtube.com/watch?v=PS5QSGAoLNw&feature=youtu.be
5. https://www.youtube.com/watch?v=8AtrYcqO5ho&feature=youtu.be
6. 4.https://www.youtube.com/watch?v=iJmJhxIsmb8&feature=youtu.be
7. https://www.youtube.com/watch?v=Gd1sVPOYzuk&feature=youtu.be
8. https://www.youtube.com/watch?v=LtGliUam-_U&feature=youtu.be
9. https://www.youtube.com/watch?v=sVrSx4zt8ho&feature=youtu.be

Course Code	Course Name	Category	L	P	T	Credit
	CLOUD PLATFORMS IN INDUSTRY	E	3			3

Preamble

- To get introduced to the cloud platforms used in the industry such as Microsoft Azure, Amazon Web Services, Google Cloud Platform (GCP).
- To learn about all the services offered by the cloud platforms.

Prerequisite

Introduction to cloud computing

Course Outcomes

On the successful completion of the course, students will be able to

Course Outcomes		Bloom's Level
CO1	Know Microsoft Azure, AWS and GCP	Remember, Understand
CO2	Use services offered by Microsoft Azure, AWS and GCP	Apply, Analyze, Evaluate, Create
CO3	Know difference between Azure, AWS and GCP	Understand, Analyze, Evaluate

Mapping with Programme Outcomes

COs	PSO1	PSO2	PSO3	PSO4	PSO5	PSO6	PSO7
CO1	S	S	M	M	S	M	L
CO2	M	M	S	S	S	M	L
CO3	S	S	M	S	S	M	L

Assessment Pattern

Category	Continuous Internal Assessment (25)			Terminal Examination (75)
	I	II	III	
Remember	5	5	5	22
Understand	6	6	6	23
Apply	5	5	5	10
Analyze	5	5	5	10
Evaluate	2	2	2	5
Create	2	2	2	5

Syllabus

Unit 1 Microsoft Azure: The benefits of cloud computing - Microsoft Azure - Cloud service models and its types - Azure subscriptions & management groups - Azure resources and Azure Resource Manager - Azure regions and availability zones. Microsoft Azure Database: Azure Cosmos DB - Azure SQL Database - Azure Database for MySQL - Azure Virtual Machines - Azure Container and Kubernetes services - Azure App Service - Azure Functions. **(11hrs)**

Unit 2 Microsoft Azure Storage & Networking Services: Azure storage account and disk storage fundamentals - Azure files - Azure virtual networks – Azure virtual network settings – Azure VPN gateway - Azure ExpressRoute. **(8hrs)**

Unit 3 Amazon Web Services: Introduction to AWS – AWS Architecture– Management Console– Console Mobile App– AWS Account – AWS Elastic Compute Cloud – AWS WorkSpaces - Virtual Private Cloud. **(8hrs)**

Unit 4 Amazon Storage, Database and Analytic Services: Amazon S3 – AWS Relational Database Service – AWS Elastic MapReduce – AWS Data Pipeline - AWS Machine Learning. **(8hrs)**

Unit 5 Google Cloud Platform: Introduction to Google Cloud Platform – benefits and features – Google Cloud Platform Services: Compute, networking, storage, big data, Security and Identity Management, Cloud AI, IoT – Creating a Free Tier Account on GCP - Difference between Google Cloud Platform, AWS and Azure. **(10hrs)**

Total: 45 hrs

Reference Books and URLs:

1. https://www.coursera.org/learn/microsoft-azure-cloud-services/lecture/ZXcFG/what-is-azure
2. https://www.tutorialspoint.com/amazon_web_services/index.htm
3. https://www.javatpoint.com/google-cloud-platform

Course Code	Course Name	Category	L	P	T	Credit
	STORM FOR DATA ANALYSTICS	E	3			3

Preamble

- To get introduced to make a career in Big Data Analytics using Apache Storm framework.
- To create and deploying a Storm cluster in a distributed environment.

Prerequisite

Basic knowledge of Java and Linux

Course Outcomes

On the successful completion of the course, students will be able to

Course Outcomes		Bloom's Level
CO1	Know the basics of Storm	Remember, Understand
CO2	Implement Storm application programming	Apply, Analyze, Create
CO3	Create and deploying a Storm cluster in a distributed environment	Apply, Analyze, Create

Mapping with Programme Outcomes

COs	PSO1	PSO2	PSO3	PSO4	PSO5	PSO6	PSO7
CO1	S	S	M	M	M	L	L
CO2	M	M	S	S	S	L	L
CO2	M	M	S	S	S	L	L

Assessment Pattern

Category	Continuous Internal Assessment (25)			Terminal Examination (75)
	I	II	III	
Remember	5	5	5	22
Understand	8	8	8	23
Apply	5	5	5	10
Analyze	5	5	5	10
Evaluate				
Create	2	2	2	5

Syllabus

UNIT I-Fundamentals of JAVA: Introduction to JAVA: Java Features – JDK – JVM - Objects and Class in Java – Data types -Arrays –Overriding methods – Abstract Class - Inheritance – Interface – Packages- Exception Handling. **(8hrs)**

UNIT II- Introduction to Storm: Introduction - Apache Storm vs Hadoop – Use Cases of Apache Storm - Apache Storm – Benefits – Core Concepts - Topology - Tasks - Workers - Stream Grouping - cluster architecture -Distributed messaging system – Work Flow - Thrift protocol. **(9hrs)**

UNIT III- Installation: Verifying Java Installation - Zoo Keeper Framework Installation - Apache storm Framework Installation - Working Examples - Scenario – Mobile Call Log Analyzer - Spout Creation - Bolt Creation - Call log Creator Bolt - Call log Counter Bolt - Creating Topology - Local Cluster - Building and Running the Application - Non-JVM **(10hrs)**

UNIT IV-Apache Trident: Trident Topology - Trident Tuples - Trident Spout - Trident Operations - State Maintenance - Distributed RPC - using Trident - Working Example of Trident - Building and Running the Application - Apache Storm In Twitter - Hashtag Reader Bolt - Hashtag Counter Bolt - Submitting a Topology - Building and Running the application. **(10hrs)**

UNIT V-Apache Storm and Applications: Spout Creation - Bolt Creation - Submitting a Topology - Building and Running the Application - Storm Applications - Klout - The Weather Channel - Telecom Industry- Yahoo Finance. **(8hrs)**

Total: 45 hrs

Reference books and URLs:

1. "JAVA: The Complete Reference", Herbert Schildt, Seventh Edition.
2. "Mastering Apache Storm", Ankit Jain.
3. "Storm Real-Time Processing Cookbook", Quinton Anderson, Packt Publishing.

Course Code	Course Name	Category	L	P	T	Credit
	SPARK FOR DATA ANALYSTICS	E	3			3

Preamble

- To get introduced to the fundamentals of Spark**,** the technology that is revolutionizing the analytics and big data world.

Prerequisite

Introduction to Data Analytics

Course Outcomes

On the successful completion of the course, students will be able to

Course Outcomes		Bloom's Level
CO1	Know the basics of Spark	Remember, Understand
CO2	Develop Spark application programs	Apply, Analyze, Create
CO3	Configure, monitor and tune Spark application	Apply, Analyze, Create

Mapping with Programme Outcomes

COs	PSO1	PSO2	PSO3	PSO4	PSO5	PSO6	PSO7
CO1	S	S	M	M	S	L	L
CO2	M	M	S	S	S	L	L
CO3	M	M	S	S	S	L	L

Assessment Pattern

Category	Continuous Internal Assessment (25)			Terminal Examination (75)
	I	II	III	
Remember	5	5	5	22
Understand	8	8	8	23
Apply	5	5	5	10
Analyze	5	5	5	10
Evaluate				
Create	2	2	2	10

Syllabus

Unit I-Introduction to Spark: Spark introduction-apache spark-evolution of apache spark-features of apache spark-spark built on hadoop-components of spark. **(7hrs)**

Unit II- SPARK RDD: Resilient Distributed Datasets-Data Sharing is Slow in MapReduce-Iterative Operations on Map Reduce-Interactive Operations on Map Reduce Data -Sharing using Spark RDD-Iterative Operations on Spark RDD-Interactive Operations on Spark RDD. **(10hrs)**

Unit III- Spark Application Programming: Purpose and usage of the Spark Context-Initialize Spark with the various programming languages-Describe and run some Spark examples-Pass functions to Spark-Create and run a Spark standalone application-Submit applications to the cluster. **(10hrs)**

Unit IV-SPARK – Core Programming: Spark Shell - RDD - Transformations -Actions - Programming with RDD-UN Persist the Storage. **(8hrs)**

Unit V- Spark configuration, monitoring and tuning: Understand components of the Spark cluster-Configure Spark to modify the Spark properties- environmental variables- or logging properties-Monitor Spark using the web UIs- metrics and external instrumentation-Understand performance tuning considerations. **(10hrs)**

Total: 45 hrs

Reference books and URLs:

1. "LearningSpark", Holden Karau, Andy Konwinski, Patrick Wendell, and Matei Zaharia.
2. "Spark: The Definitive Guide", Matei Zaharia, Bill Chambers, O'Reilly Media, Inc, 2018, ISBN:9781491912201.