College of Arts, Science & Commerce

RISE WITH EDUCATION

Sion (West), Mumbai – 400022.

(Autonomous)

Faculty: Science

Program: M.Sc.

Subject: DATA SCIENCE

PART II

Academic Year: 2021 – 2022

Credit Based Semester and Grading Syllabi approved by Board of Studies in Data science to be brought into effect from July 2021

M.Sc(Data Science)

SEMESTER – III			SEMESTER – IV		
Subject Code	**Subject Name**	**Credits**	**Subject Code**	**Subject Name**	**Credits**
SIPSDS31	Big Data Analytics	4	SIPSDS41	Deep Learning	4
SIPSDS32	Linear Algebra	4	SIPSDS42	Web and Social Network Data Analytics	4
SIPSDS33	Data Visualization	4	SIPSDSP41	Deep Learning practical	2
SIPSDS34	Data Storage and Management	4	SIPSDSP42	Web and Social Network Data Analytics	2
SIPSDSP31	Big Data Analytics Practical	2	SIPSDS43	Internship	6
SIPSDSP32	Linear Algebra Practical	2			
SIPSDSP33	Data visualization Practical	2			
SIPSDSP34	Data Storage and Management Practical	2			
SIPSDSP35	Research Project - Proposal	2	SIPSDS44	Research Project - Implementation	4
Total Credits		**26**	**Total Credits**		**22**

Big Data analytics

Learning Objective: The main goal of this course is to help students learn, understand, and practice big data analytics approaches, which include the conceptualization and summarization of big data and machine learning, and big data computing technologies.

Learning Outcome:

- Ability to identify the characteristics of datasets and compare the trivial data and big data for various applications.
- Ability to solve problems associated with batch learning and online learning, and the big data characteristics such as high dimensionality, dynamically growing data and in particular scalability issues.

Theory Component:

M. Sc (Data Science)	**Semester – III – SIPSDS31**
Course Name	**Big Data analytics**
Periods per week (1 Period is 60 minutes)	**4**
Credits (Theory + Internals)	**4**

Unit	Contents	No. of Lectures
I	**Introduction:** Introduction to Big Data, Big Data Characteristics, Types of Big Data, Traditional Versus Big Data Approach, Technologies Available for Big Data, Infrastructure for Big Data, Use of Data Analytics, Big Data Challenges, Desired Properties of a Big Data System, Case Study of Big Data Solutions	**12**
II	**Analytical Theory and Methods:** Clustering and Associated Algorithms, Association Rules, Apriori Algorithm, Candidate Rules, Applications of Association Rules, Validation and Testing, Diagnostics, Regression, Linear Regression, Logistic Regression, Additional Regression Models.	**12**
III	Classification, Decision Trees, Naïve Bayes, Diagnostics of Classifiers, Additional Classification Methods, Time Series Analysis, Box Jenkins methodology, ARIMA Model, Additional methods. Text Analysis, Steps, Text Analysis Example, Collecting Raw Text, Representing Text, Term Frequency-Inverse Document Frequency (TFIDF), Categorizing Documents by Topics, Determining Sentiments	**12**

IV	**Hadoop:** Introduction, What is Hadoop?, Core Hadoop Components, Operating System for Big Data, Concepts, Hadoop Architecture, Hadoop Ecosystem, Hive, , Hadoop Limitations , Recommendation Systems.	**12**
V	**NoSQL:** What is NoSQL?, NoSQL Business Drivers, NoSQL Case Studies, NoSQL Data Architectural Patterns, Variations of NoSQL Architectural Patterns, Using NoSQL to Manage Big Data Map Reduce: MapReduce and The New Software Stack, MapReduce, Algorithms Using MapReduce	**12**

Books and References

Sr. No.	Title	Author/s	Publisher	Edition	Year
1	Big Data Analytics	Radha Shankarmani	Wiley	Second	2016
2	Big Data and Analytics	Subhashini Chellappan Seema Acharya	Wiley	First	2015
3	Big Data Analytics with R and Hadoop	Vignesh Prajapati	Packt	First	2013
4	Practical Big data Analytics	Nataraj Dasgupta	Pack	First	2018
5	Big Data Analytics	Anuradha Bhatia			

Linear Algebra

Learning Objectives:
To offer the learner the relevant linear algebra concepts through Data science applications.

Learning Outcomes:

- Appreciate the relevance of linear algebra in the field of computer science.
- Understand the concepts through program implementation
- Instill a computational thinking while learning linear algebra and linear programming.
- Linear Programming (LP), also known as linear optimization is a mathematical programming technique to obtain the best result or outcome.

Theory Component:

M. Sc (Data Science)	**Semester – III – SIPSDS32**
Course Name	**Linear Algebra**
Periods per week (1 Period is 60 minutes)	**4**
Credits (Theory + Internals)	**4**

Unit	Contents	No. of Lectures
I	**Field**: Introduction to complex numbers, numbers in Python , Abstracting over fields, Playing with GF(2), Vector Space: Vectors are functions, Vector addition, Scalar-vector multiplication, Combining vector addition and scalar multiplication, Dictionary-based representations of vectors, Dot-product, Solving a triangular system of linear equations. Linear combination, Span, The geometry of sets of vectors, Vector spaces, Linear systems, homogeneous and otherwise	**12**
II	**Matrix**: Matrices as vectors, Transpose, Matrix-vector and vector-matrix multiplication in terms of linear combinations, Matrix-vector multiplication in terms of dot-products **Null space**: General description, Computing sparse matrix-vector product, Linear functions, Matrix-matrix multiplication, Inner product and outer product, From function inverse to matrix inverse **Basis**: Coordinate systems, Two greedy algorithms for finding a set of generators, Minimum Spanning Forest and GF(2), Linear dependence, Basis ,Unique representation, Change of basis, first look, Computational problems involving finding a basis	**12**

III	**Dimension:** Dimension and rank, Direct sum, Dimension and linear functions, The annihilator **Linear transformations :** properties, matrix of a linear transformation, change of basis, range and kernel, rank and nullity, Rank, Nullity theorem **Gaussian elimination**: Echelon form, Gaussian elimination over GF(2), Solving a matrix-vector equation using Gaussian elimination, Finding a basis for the null space, Factoring integers,	**12**
IV	**Inner Product**: The inner product for vectors over the reals, Orthogonality, **Orthogonalization**: Projection orthogonal to multiple vectors, Projecting orthogonal to mutually orthogonal vectors, Building an orthogonal set of generators, Orthogonal complement	**12**
V	**Eigenvector**: Modeling discrete dynamic processes, Diagonalization of the Fibonacci matrix, Eigenvalues and eigenvectors, Coordinate representation in terms of eigenvectors, The Internet worm, Existence of eigenvalues, Markov chains, Modeling a web surfer: Page Rank. **Linear Algebra**: Applications, vectorized code, image recognition, dimensionality reduction.	**12**

Books and References

Sr. No.	Title	Author/s	Publisher	Edition	Year
1	Coding the Matrix Linear Algebra through Applications to Computer Science	PHILIP N. KLEIN	Newtonian Press	1	2013
2	Linear Algebra and Its Applications	Gilbert Strang	Cengage Learning	4th	2007
3	Linear Algebra and Its Applications	David C Lay	Pearson Education India	3rd	2002
4	Linear Algebra and Probability for Computer Science Applications	Ernest Davis, A K Peters	A K Peters	1	2012
5	Operation research	SD Sharama	Kedarnath	2017	2012

Data Visualization

Learning Objectives:

- To apply the functionality of the various data visualization tools and techniques
- To understand visual perception, visual representation of data

- To understand and apply various classification and prediction techniques using tools.
- To study and apply Visualization of groups, trees, graphs, clusters, networks on data set.
- To understand Mining of Object, Spatial, Multimedia, Text and Web Data.

Learning Outcomes: -

- Analyze the visual representation of data on time series and statistical data.
- Apply visual mapping, visual analytics,
- Design of visualization applications
- Analyze the Classification of visualization systems and metaphorical visualization.

Theory Component:

M. Sc (Data Science)	**Semester – III – SIPSDS33**
Course Name	**Data Visualization**
Periods per week (1 Period is 60 minutes)	**4**
Credits (Theory + Internals)	**4**

Unit	Contents	No. of Lectures
I	Introduction of visual perception, visual representation of data, Gestalt principles, information overloads, Design principles Categorical, time series, and statistical data graphics.	**12**
II	Creating visual representations, visualization reference model, visual mapping, visual analytics, Design of visualization applications.	**12**
III	Classification of visualization systems, Interaction and visualization techniques misleading, Visualization of one, two and multi-dimensional data, text and text documents.	**12**
IV	Visualization of groups, trees, graphs, clusters, networks, software, Metaphorical visualization	**12**
V	Visualization of volumetric data, vector fields, processes and simulations, Visualization of maps, geographic information, GIS systems, collaborative visualizations, evaluating visualizations	**12**

Books and References

Sr. No.	Title	Author/s	Publisher	Edition	Year

1	Interactive Data Visualization: Foundations, Techniques, and Applications.	Ward, Grinstein Keim	A K Peters/CRC Press	Second	2015
2	The Visual Display of Quantitative Information	E. Tufte	Graphics Press	Second	2001

Data Storage and Management

Learning Objectives: -

- Understand the types of storage systems.
- Utilize redundant array of independent disks (RAID) technologies effectively
- Understanding the positioning of data at various level of memory hierarchy.
- Learning Distributed data base system, Mango DB, Storage architecture, AN.

Learning Outcomes: -

- Analyze the data center requirements for a business setup and apply the right information cycle
- Apply the best storage configuration in distributed environment.
- Select the best techniques for facilitation backup and recovery of lost or corrupted data
- Design, analyze storage systems and select an optimal storage network
- Design and compare cloud storage setup for efficient business transaction setup

Theory Component:

M. Sc (Data Science)	**Semester – III – SIPSDS34**
Course Name	**Data Storage and Management**
Periods per week (1 Period is 60 minutes)	**4**
Credits (Theory + Internals)	**4**

Unit	Contents	No. of Lectures
I	Storage Media and Technologies – Magnetic, Optical and Semiconductor Media, Techniques for read/write Operations, Issues and Limitations.	**12**
II	Usage and Access – Positioning in the Memory Hierarchy, Hardware and Software Design for Access, Performance issues. Distributed Database Patterns— Distributed Relational Databases- Non-relational Distributed Databases- MongoDB - Sharing and Replication- HBase-	**12**
III	Cassandra Consistency Models— Types of Consistency- Consistency MongoDB- HBase Consistency- Cassandra Consistency. Large Storages – Hard Disks, Networked Attached Storage, Scalability issues, Networking issues.	**12**
IV	Storage Architecture - Storage Partitioning, Storage System Design, Caching, Legacy Systems.	**12**
V	Storage Area Networks – Hardware and Software Components, Storage Clusters/Grids. Storage QoS–Performance, Reliability, and Security issues, storage appliances. Network and web security: Network Security: Network Concepts, Threats in Networks, Network Security Controls. Web Security Requirements,	**12**

	Secure Socket Layer (SSL), Transport Layer Security (TLS), Secure Electronic Transaction (SET).	

Books and References

Sr. No.	Title	Author/s	Publisher	Edition	Year
1	The Complete Guide to Data Storage Technologies for Network-centric Computing	Franklyn E. Dailey Jr.	Computer Technology Research Corporation	First	1998.
2	Data Storage Networking	Nigel Poulton	Sybex	First	2014

Practical Component: (SEMESTER III)

M. Sc (Data Science)	**Semester – III – SIPSDSP31**
Course Name	**Big Data analytics**
Periods per week (1 Period is 60 minutes)	**4**
Credits	**2**

List of Practical:

1	Installation of HADOOP
2	Implement the following file management tasks in Hadoop System (HDFS): Adding files and directories, Retrieving files, Deleting files
3	Basic CRUD operations in MongoDB
4	To understand the overall programming architecture using Map Reduce API and implement programs related to MapReduce
5	Implement clustering and associated algorithms
6	Implement Linear Regression
7	Implement Bloom Filters for filter on Stream Data
8	Implement Time Series
9	Creating the HDFS tables and loading them in Hive and learn joining of tables in Hive
10	To perform NoSQL database using mongodb to create, update and insert.
11	Implement a simple recommender system

M. Sc (Data Science)	**Semester – III – SIPSDSP32**
Course Name	**Linear Algebra Practical**
Periods per week (1 Period is 60 minutes)	**4**
Credits	**2**

List of Practical: Implement using R/Python Programming.

1. Write a program which demonstrates the following:
   a. Addition of two complex numbers
   b. Displaying the conjugate of a complex number
   c. Plotting a set of complex numbers
   d. Creating a new plot by rotating the given number by a degree 90, 180, 270 degrees and also by scaling by a number a=1/2, a=1/3, a=2 etc.
2. Write a program to do the following:
   a. Enter two distinct faces as vectors u and v.
   b. Find a new face as a linear combination of u and v i.e. au+bv for a and b in R.
   c. Find the average face of the original faces.
3. Write a program to do the following:
   a. Enter a vector u as a n-list
   b. Enter another vector v as a n-list
   c. Find the vector au+bv for different values of a and b
   d. Find the dot product of u and v
4. Write a program to do the following:
   a. Enter an r by c matrix M (r and c being positive integers)
   b. Display M in matrix format
   c. Display the rows and columns of the matrix M
   d. Find the scalar multiplication of M for a given scalar.
   e. Find the transpose of the matrix M.
5. Write a program to do the following:
   a. Find the vector –matrix multiplication of a r by c matrix M with an c-vector u.
   b. Find the matrix-matrix product of M with a c by p matrix N.
6. Write a program to enter a matrix and check if it is invertible. If the inverse exists, find the inverse.
7. Write a program to convert a matrix into its row echelon form
8. Write a program to find Eigen values and vectors
9. Write a program to implement gaussion elimination method.
10. Write a program to implement concepts of orthogonalization.

M. Sc (Data Science)	**Semester – III – SIPSDSP33**
Course Name	**Data Visualization**
Periods per week (1 Period is 60 minutes)	**4**
Credits	**2**

List of Practical:

1) Demonstrate a nonobvious insight gleaned from the data, or to make a particular point. You can stick with a single chart or other type of visualization, or you can use multiple displays that together tell a story. To create your visualization(s) you can use the simple tools - spreadsheet graphing tools (e.g., Google Sheets, Excel), Tableau tool, or any other method of creating a visual point or story from the data.
2) Demonstrate Time series and statistical data graphics using visualization tool.
3) Generating visualizations of map-based data.
4) Finding data There are many sources of freely downloadable data. Locate a relevant data set online.
   Here are some places to start:
   • LION: http://www.nyc.gov/html/dcp/html/bytes/applbyte.shtml
   •Newman Library: http://guides.newman.baruch.cuny.edu/nyc_data
   • NYC Open Data: http://data.cityofnewyork.us
   • Wiki: https://wiki.gephi.org/index.php/Datasets
   Demonstrate temporal component, showing change over time.
5) Demonstrate visualization of one, two and multi-dimensional data.
6) Visualizing tenure, monthly charges, total charges, and other individual columns using a scatter plot
7) Demonstrate visualization of text and text documents.
8) Create a Map view with appropriate data set using tableau.
9) Demonstrate Metaphorical visualization.
10) Demonstrate visualization of groups, trees, graphs, clusters.
11) Demonstrate collaborative visualization

M. Sc (Data Science)	**Semester – III – SIPSDSP34**
Course Name	**Data Storage and Management**
Periods per week (1 Period is 60 minutes)	**4**
Credits	**2**

List of Practical:

1. Demonstrate the Read/Write time of data from various storage devices (Pend Derive, HDD, CD/DVD)
2. Demonstrate the Usage and Access of data positioned at various level of memory hierarchy.
3. Build an application on private cloud.
4. Demonstrate any Cloud data storage Monitoring tool.
5. Implement FOSS-Cloud Functionality - VDI (Virtual Desktop Infrastructure)
6. Demonstrate Distributed Relational Databases
7. Demonstrate Non-relational Distributed Databases
8. Demonstrate parallel data base in peer-to-peer environment.
9. Implement FOSS-Cloud Functionality - VSI Software as a Service (SaaS).
10. Explore Working of the following with Virtual Machines-
    a. VM Lifecycle
    b. Creating VMs
    c. Accessing VMs
    d. Assigning VMs to Hosts
11. Explore Service Offerings, Disk Offerings, Network Offerings and Templates – In open source Cloud technology

Deep Learning

Learning Objectives:-

- To know importance of deep learning
- To acquire knowledge on the basics of neural networks.
- To implement neural networks using computational tools for variety of problems.
- To explore various deep learning algorithms

Learning Outcomes: -

- Develop algorithms simulating human brain.
- Analyze ANN learning and memory-based learning
- Explore the essentials of Deep Learning and Deep Network architectures.
- Define, train and use a Deep Neural Network for solving real world problems that require artificial Intelligence based solutions.
- Use deep learning methodology in real world application

Theory Component:

M. Sc (Data Science)	**Semester – IV – SIPSDS41**
Course Name	**Deep Learning**
Periods per week (1 Period is 60 minutes)	**4**
Credits (Theory + Internals)	**4**

Unit	Contents	No. of Lectures
I	Introduction: Feed forward neural networks. Gradient descent and the back propagation algorithm. Unit saturation, aka the vanishing gradient problem, and ways to mitigate it. RelU Heuristics for avoiding bad local minima. Heuristics for faster training. Nestors accelerated gradient descent. Regularization. Dropout.	**12**

II	Convolution Neural Network: Architectures, convolution / pooling layers. Applications of Deep Learning to Computer Vision : Image segmentation, object detection, automatic image captioning, Image generation with Generative adversarial networks, video to text with LSTM models. Attention models for computer vision tasks.	**12**
III	Recurrent Neural Networks: LSTM, GRU, Encoder Decoder architectures. Recent Reseearch in NLP using Deep Learning: Factoid Question Asnwering, similar question detection, Dialogue topic tracking, Neural Summarization, Smart Reply	**12**
IV	Deep Unsupervised Learning: Auto encoders (standard, sparse, denoising, contractive, etc), Variational Auto encoders, Adversarial Generative Networks, Auto encoder and DBM.	**12**
V	Applications of Deep Learning to Computer Vision Image segmentation, object detection, automatic image captioning, Image generation with Generative adversarial networks, and video to text with LSTM models. Attention models for computer vision tasks.	**12**

Books and References

Sr. No.	Title	Author/s	Publisher	Edition	Year
1	Deep Learning	Ian Goodfellow, Yoshua Bengio, Aaron Courville	MIT Press		2016
2	Neural Networks and Deep Learning	Michael Nielsen	Determination Press		2015
3	Deep Learning with Python	Francois Chollet	Manning Publications	First	2017

Web and Social Network Data Analytics

Learning Objective:

- To introduce the concepts of Web and Information Retrieval and Web Mining in Social Network.
- To study the basic concepts of Social Network Analysis.
- To interpret Social networks through mathematical representation.
- To analyze relations, descriptive measures and models to overview research questions related to Social Networks.

Learning Outcome:

The students will be able to:-

- Choose and analyze various Information Retrieval Models and in turn will be able to develop Information Retrieval Systems
- Gather relevant network data, and some of the associated questions and problems.
- To build various applications based on Social Network platform.

Theory Component:

M. Sc (Data Science)	**Semester – IV – SIPSDS42**
Course Name	**Web and Social Network Data Analytics**
Periods per week (1 Period is 60 minutes)	**4**
Credits (Theory + Internals)	**4**

Unit	Contents	No. of Lectures
I	**Information Retrieval and Web Search:** Basic Concepts, Information Retrieval Models, Text and Web Page Pre-Processing, Inverted Index and Its Compression, Latent Semantic Indexing, Web Search, Meta-Search, Web Spamming	**12**

II	**Social Network Analysis**: Co-Citation and Bibliographic Coupling, PageRank, HITS Algorithm, Community Discovery **Web Crawling**: A Basic Crawler Algorithm, Implementation Issues, Universal Crawlers, Topical Crawlers, Crawler Ethics and Conflicts	**12**
III	**Structured Data Extraction: Wrapper Generation:** Preliminaries, Wrapper Induction, Instance-Based Wrapper Learning, Automatic Wrapper Generation, String Matching and Tree Matching, Multiple Alignment, Flat Data Records, Nested Data Records, Extraction Based on Multiple Pages	**12**
IV	**Opinion Mining and Sentiment Analysis:** The Problem of Opinion Mining, Document Sentiment Classification, Sentence Subjectivity and Sentiment Classification, Aspect-Based Opinion Mining, Mining Comparative Opinions, Opinion Search and Retrieval, Opinion Spam Detection	**12**
V	**Web Usage Mining:** Data Collection and Pre-Processing, Data Modeling, Discovery and Analysis of Web Usage Patterns, Recommender Systems and Collaborative Filtering, Query Log Mining	**12**

Books and References

Sr. No.	Title	Author/s	Publisher	Edition	Year
1	Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data	Bing Liu	Springer	Second	2011
2	Mining the Social Web	Matthew A. Russell and Mikhail Klassen	O'Reilly	Third	2019
3	Analyzing Social Networks.	Stephen P. Borgatti	SAGE	First	2013

Practical Component: (SEMESTER IV)

M. Sc (Data Science)	**Semester – IV – SIPSDSP41**
Course Name	**Deep Learning Practical**
Periods per week (1 Period is 60 minutes)	**4**
Credits	**2**

List of Practical:

1) Demonstrate Gradient descent and the back-propagation algorithm.
2) Implement RelU Heuristics for avoiding bad local minima.
3) Demonstrate Image segmentation, object detection.
4) Demonstrate automatic image captioning, Image generation with Generative adversarial networks.
5) Demonstrate video to text with LSTM model.
6) Demonstrate Neural Summarization using NLP
7) Demonstrate similar question detection using NLP.
8) Demonstrate Auto encoders using Deep Unsupervised Learning.
9) Demonstrate Variational Auto encoders.
10) Demonstrate Dialogue topic tracking.

M. Sc (Data Science)	**Semester – IV – SIPSDSP42**
Course Name	**Web and Social Network Data Analytics**
Periods per week (1 Period is 60 minutes)	**4**
Credits	**2**

List of Practical: (To be implemented using any of the web mining tools)

1	Page Rank Algorithm
2	Weighted Page Rank Algorithm
3	HITS Algorithm
4	Crawler Algorithms
5	Structured Data Extraction through Wrapper Generation
6	Opinion Search and Retrieval
7	Sentiment Analysis
8	Web Content Mining
9	Web Structure Mining
10	Case Studies: Google, Facebook, Twitter, Instagram