Syllabus for

Two Years M.Sc. Program

in

DATA SCIENCE

Under

BERHAMPUR UNIVERSITY

BERHAMPUR – 760 007, ODISHA

2023

M.Sc. in Data Science

SEM	COURSE CODE	COURSE NAME	CREDIT	MARKS		TOTAL
				Mid Sem	End Sem	
I	DS C101	Mathematical Foundation for Data Science	4	20	80	100
	DS C102	Introduction to Probability and Statistics	4	20	80	100
	DS C103	Not only SQL Databases	4	20	80	100
	DS C104	Computer Networks and Cybersecurity	4	20	80	100
	DS P105	Lab I – R Programming and MySQL and NoSQL (**Practical**)	4	20	80	100
II	DS C201	Foundation of Data Science and Analytics	4	20	80	100
	DS C202	Cloud Computing	4	20	80	100
	DS C203	Business Analytics	4	20	80	100
	DS C204	Internet of Things	4	20	80	100
	DS P205	Lab II – IoT and Cloud Computing (**Practical**)	4	20	80	100
	DS VAC206	**Data Analysis with Power BI**	**Non-Credit Course**			
III	DS C301	Machine Learning	4	20	80	100
	DS C302	Big Data Analytics	4	20	80	100
	DS CBCS *	Data Analytics	4	20	80	100
	ELECTIVES					
	DS E303	Predictive Analytics	4	20	80	100
	DS E304	Decision Management Systems	4	20	80	100
	DS E305	Text Analytics	4	20	80	100
	DS P306	Lab III – Python Programming and Machine Learning (**Practical**)	4	20	80	100
	DS VAC 307	**Programming with NumPy and Pandas**	**Non-Credit Course**			
IV	DS C401	Social Media Analytics	4	20	80	100
	DS C402	Time Series Analysis and Forecasting	4	20	80	100
	ELECTIVES					
	DS E403	Healthcare Analytics	4	20	80	100
	DS E404	Sentiment Analytics	4	20	80	100
	DS E405	Image and Video Analytics	4	20	80	100
	DS PR406	Project Work / Dissertation	8			200
					TOTAL	**2000**

Note: C – Core Course, E – Elective Course, P – Practical, PR - Project

A student has to opt. for one elective course in a semester.

*** CBCS Course in 3rd sem to be offered by Computer Science Department for others.**

Semester – I

DS C101	Mathematical Foundation for Data Science	4	20	80	100
DS C102	Introduction to Probability and Statistics	4	20	80	100
DS C103	Not only SQL Databases	4	20	80	100
DS C104	Computer Networks and Cybersecurity	4	20	80	100
DS P105	Lab I – R Programming and MySQL and NoSQL (PRACTICAL)	4	20	80	100

SEMESTER - I

COURSE CODE: DS C101	Mathematical Foundation for Data Science
CORE COURSE	4 CREDITS

Unit-1

Vector Spaces: R^n and C^n, lists, F^{n}and digression on Fields, Definition of Vector spaces, Subspaces, sums of Subspaces, Direct Sums, Span and Linear Independence, bases, dimension.

Unit-2

Definition of Linear Maps - Algebraic Operations on - Null spaces and Injectivity - Range and Subjectivity -Fundamental Theorems of Linear Maps - Representing a Linear Map by a Matrix - Invertible Linear Maps -Isomorphic Vector spaces - Linear Map as Matrix Multiplication - Operators - Products of Vector Spaces -Product of Direct Sum - Quotients of Vector spaces (Basic only).

Unit-3

Eigenvalues and Eigenvectors - Eigenvectors and Upper Triangular matrices – Eigenspaces and Diagonal Matrices - Inequalities on Linear Spaces - Norms on Linear Spaces - Inner products - Orthogonality - Unitary and Orthogonal Matrices - Norms for matrices, SVD, Least Square Solution, Moore-Penrose Inverse.

Unit-4

Functions of Several Variables - Limits and continuity in Higher Dimensions – Partial Derivatives - The Chain Rule - Directional Derivative and Gradient vectors - Tangent Planes and Differentials - Extreme Values and Saddle Points - Lagrange Multipliers.

Graphs - subgraphs - factors - Paths - cycles - connectedness -trees - Euler tours - Hamiltonian cycles - Planar Graphs - Digraphs.

TEXT BOOKS

1) S. Axler, Linear algebra done right, Springer, 2017.
2) Eldén Lars, Matrix methods in data mining and pattern recognition, Society for Industrial and Applied Mathematics, 2007.
3) M. D. Weir, J. Hass, and G. B. Thomas, Thomas' calculus. Pearson, 2016.
4) D. Jungnickel, Graphs, networks and algorithms. Springer, 2014.
5) Gilbert Strang, Linear Algebra and Its Applications, 4th edition.

REFERENCES

1) E. Davis, Linear algebra and probability for computer science applications, CRC Press, 2012.
2) J. V. Kepner and J. R. Gilbert, Graph algorithms in the language of linear algebra, Society for Industrial and Applied Mathematics, 2011.
3) D. A. Simovici, Linear algebra tools for data mining, World Scientific Publishing, 2012.
4) P. N. Klein, Coding the matrix: linear algebra through applications to computer science, Newtonian Press, 2015.
5) J. Patterson and A. Gibson, Deep learning: a practitioner's approach. O'Reilly Media, 2017. 6. S. Sra, S. Nowozin, and S. J. Wright, Optimization for machine learning. MIT Press, 2012.

COURSE CODE: DS C102	Introduction to Probability and Statistics
CORE COURSE	4 CREDITS

Unit-1

Introduction to Statistics: Data Collections and Descriptive Statistics, Inferential Statistics and Probability Models, Populations and Samples.

Descriptive Statistics: Describing and Summarizing Data Sets, Chebyshev's Inequality, Normal and Paired Data Sets, Sample Correlation Coefficient Problems.

Elements of Probability: Sample Space and Events, Venn Diagrams and Algebra of Events, Axioms of Probability, Sample Spaces Having Equally Likely Outcomes, Conditional Probability, Bayes' Formula, Independent Events.

Unit-2

Random Variables and Expectations: Random Variables, Types of Random Variables, Jointly distributed Random Variables, Expectation, Properties of the expected value, Variance, Covariance and Variance of Sums of Random Variables, Moment Generating Functions. Chebyshev's Inequality and the Weak Law of Large Numbers Problems.

Special Random Variables: Bernoulli and Binomial Distribution Function, Poisson Random Variables, Hypergeometric Random Variable, Uniform Random Variable, Normal Random Variables, Exponential Random Variables, Poisson Process, Gamma Distribution, Chi-Square Distribution, t-Distribution, f-Distribution, Logistics Distribution (Basics only).

Unit-3

Distributions of Sampling Statistics: The Sample Mean, Central Limit Theorem, Sample Variance, Sampling Distributions from a Normal Population, Sampling from a Finite Population.

Parameter Estimation: Maximum Likelihood Estimators, Internal Estimates, Estimating the difference in Means of Two Normal Populations, Approximate Confidence Interval for the Mean of a Bernoulli Random Variable.

Unit-4

Hypothesis Testing: Significance levels, Tests Concerning the Mean of a Normal Population, Testing the Equality of Means of Two Normal Populations, Hypothesis Tests in Bernoulli Populations, Introduction to Parametric and Non-parametric Text: t, z, χ^2 ANNOVA.

Co-relation and Regression: Introduction, Types of Correlation, Least Squares Estimators of the Regression Parameters, The Coefficient of Determination and the Sample Correlation Coefficient, Scatter diagram method, Kerl Parsons Method, Rank Correlation, Analysis of Residual, Transforming to Linearity, Weighted Least Squares, Polynomial Regression, Multiple Linear Regression, Predicting Future Responses, Logistic Regression Models for Binary Output Data.

TEXT BOOKS:

1) Introduction to Probability and Statistics for Engineers and Scientists by Sheldon M. Ross, Elsevier. Chapter 1-9.
2) Introduction to Probability and Statistics Using R by G. Jay Kerns.

COURSE CODE: DS C103	NoSQL Databases
CORE COURSE	4 CREDITS

Unit-1

DBMS: Terminologies, components, roles, advantages and disadvantages – Database architectures: teleprocessing, file server, 2-tier, 3-tier, N-tier, middleware and Transaction processing monitor – Software components of DBMS and Database Manager – Data modeling using ER diagram: Entity, relationship, attributes, keys, strong and weak entities, attributes on relationships, relationship types, cardinality and participation.

Unit-2

SQL Statements: SELECT, WHERE, ORDERBY, GROUPBY and HAVING clauses - Sub Queries – ANY and ALL – JOIN – inner and outer joins – EXISTS and NON EXISTS – UNION, INTERSECT and EXCEPT– Updating databases: INSERT, UPDATE and DELETE – SQL data types – Creating, altering and removing tables – Indexes and views: CREATE and REMOVE.

Unit-3

Why NoSQL – Value of Relational Database – Emergence of NoSQL – Aggregate data models – More details on data models: Relationships, Graphs DB, Schemaless DB, Materialized views – Distribution models: Single server, shrading, replication – Consistency: Update, read, relax consistency.

Unit-4

Key value databases: What is Key Value store, Features of Key value DB, Suitable use cases, When not to use it – Document databases: Definition, features, Suitable use cases, when not to use – Column family stores: Definition, features, suitable use cases, when not to use – Graph databases: Definition, features, use case, when not to use – Schema migration – Polyglot persistence - Beyond NoSQL –Choosing your database

Introduction to MongoDB: Document – Collection – Database - Datatypes – Creating, deleting, updating documents – Querying – Indexing – Aggregation: Pipeline, Aggregation commands –Application design.

TEXT BOOKS

1) Thomas M. Connolly and Carolyn E. Begg. Database Systems: “*A Practical Approach to Design, Implementation, and Management”,* 6th Edition, Pearson, 2015.
2) Pramod J. Sadalage; Martin Fowler. *NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence*. Addison-Wesley. 2012 ISBN: 0321826620.
3) Kristina Chodorow, MongoDB: The Definitive Guide, 2ed, Oreilly Publishers.

REFERENCES

1) Eric Redmond; Jim R. Wilson. *Seven Databases in Seven Weeks: A Guide to Modern Databases and the NoSQL Movement*. Pragmatic Bookshelf. 2012. ISBN: 1934356921

COURSE CODE: DS C104	Computer Networks and Cybersecurity
CORE COURSE	4 CREDITS

Unit-1

Introduction to Data Communications and Network Models: Protocols and Standards, Layers in OSI Models, Analog and Digital Signals, Transmission Modes, Transmission Impairment, Data Rate Limits, Performance, Digital Transmission, Network Devices & Drivers: Router, Modem, Repeater, Hub, Switch, Bridge (fundamental concepts only).

Transmission Media: Guided Media, Unguided Media, Switching Techniques: Packet Switching, Circuit Switching, Datagram Networks, Virtual-Circuit Networks, and Structure of a Switch.

Unit-2

Error Detection and Correction: Checksum, CRC, Data Link Control: Framing, Flow and Error Control, Noiseless Channels, Noisy channels, (Stop and Wait ARQ, Slidding Window Protocol, Go Back N, Selective Repeat) HDLC, Point-to-Point Protocol. Access Control: TDM, CSMA/CD, and Channelization (FDMA, TDMA, and CDMA).

Unit-3

Network Layer: Logical Addressing, IPv4 Addresses, IPv6 Addresses, Virtual-Circuit Transport Layer Protocol, Sockets, Process-Process Delivery, UDP, TCP.

Application layers: DNS, SMTP, POP, FTP, HTTP, Basics of WiFi (Fundamental concepts only),

Unit-4

Introduction to Network Security: Confidentiality, Integrity, Authenticity (CIA) Introduction to Cryptography, Symmetric-Key Cryptography: Traditional Ciphers, Simple Modern Ciphers, Modern Round Ciphers, Mode of Operations. Asymmetric-key Cryptography: RSA and Diffie-Hellman.

Network Security: Security Services, Message Confidentiality, Message Integrity, Message Authentication: MAC and HMAC, Digital Signature, Key Management: Symmetric-key Distribution: KDC, Session Keys, Kerberos, Public-key Distribution: Certification Authority, X.509, PKI.

TEXT BOOKS:

1) Data Communications and Networking, Fourth Edition by Behrouza A. Forouzan, TMH.

REFERENCE BOOKS:

1) Computer Networks, A.S. Tanenbaum, 4th edition, Pearson Education.

DS P105 Lab. I: R Programming and MySQL and NoSQL(4 Credits)

Introduction to R Programming:

Introduction: Overview and History of R, Getting Help, Data Types, Subsetting, Vectorized, Operations.

Reading and Writing Data. Control Structures, Functions, lapply, tapply, split, mapply, apply. Coding Standards

Scoping Rules. Debugging Tools, Simulation, R Profiler

Data Analysis Using R: Visualization before Analysis, Dirty Data, Visualizing a Single Variables, Examining Multi Variable.

TEXT BOOKS:

1) William N. Venables and David M. Smith, An Introduction to R. 2ndEdition. Network Theory Limited. 2009.
2) Norman Matloff, The Art of R Programming -A Tour of Statistical Software Design, No Starch Press. 2011.

Practical: R Programming

1) Write a program that prints ‗Hello World' to the screen.
2) Write a program that asks the user for a number n and prints the sum of the numbers 1 to n.
3) Write a program that prints a multiplication table for numbers up to 12.
4) Write a function that returns the largest element in a list.
5) Write a function that computes the running total of a list.
6) Write a function that tests whether a string is a palindrome.
7) Implement the following sorting algorithms: Selection sort, Insertion sort, Bubble Sort.
8) Implement linear search.
9) Implement binary search.
10) Implement matrices addition, subtraction and Multiplication.

11) Other programs to visualize single variables and multiple variables such as histogram, density plot, Dotchart, Barplot, Box-and-Whisker Plot, Hexbin Plot, Scatterplot Matrix.

12) Space Vacation Histogram experiment using R function called SpaceVacation.r

13) Testing of Chebyshev's Theorem using R.

14) Using R to simulate experiments.

15) Using R to create Binomial Distributions.

16) Understanding of Normal Distributions and compare theoretical distribution to some real data.

17) Using R, empirically study how Central Limit Theorem works.

MySQL and NoSQL:

Practical/Tutorial: Database Systems Labs

Create and use the following database schema to answer the given queries.

EMPLOYEE Schema

Field	**Type**	**NULL**	**KEY**	**DEFAULT**
Eno	Char(3)	NO	PRI	NIL
Ename	Varchar(50)	NO		NIL
Job_type	Varchar(50)	NO		NIL
Manager	Char(3)	Yes	FK	NIL
Hire_date	Date	NO		NIL
Dno	Integer	YES	FK	NIL
Commission	Decimal(10,2)	YES		NIL
Salary	Decimal(7,2)	NO		NIL

DEPARTMENT Schema

Field	**Type**	**NULL KEY**	**DEFAULT**
Dno	Integer	No PRI	NULL
Dname	Varchar(50)	Yes	NULL
Location	Varchar(50)	Yes	New Delhi

MySQL Query List

1) Query to display Employee Name, Job, Hire Date, Employee Number; for each employee with the Employee Number appearing first.

2) Query to display unique Jobs from the Employee Table.

3) Query to display the Employee Name concatenated by a Job separated by a comma.

4) Query to display all the data from the Employee Table. Separate each Column by a comma and name the said column as THE_OUTPUT.

5) Query to display the Employee Name and Salary of all the employees earning more than$2850.

6) Query to display Employee Name and Department Number for the Employee No= 7900.

7) Query to display Employee Name and Salary for all employees whose salary is not in the range of $1500 and $2850.

8) Query to display Employee Name and Department No. of all the employees in Dept 10 and Dept 30 in the alphabetical order by name.

9) Query to display Name and Hire Date of every Employee who was hired in 1981.

10) Query to display Name and Job of all employees who don’t have a current Manager.

11) Query to display the Name, Salary and Commission for all the employees who earn commission.

12) Sort the data in descending order of Salary and Commission.

13) Query to display Name of all the employees where the third letter of their name is ‘A’.

14) Query to display Name of all employees either have two ‘R’s or have two ‘A’s in their name and are either in Dept No = 30 or their Mangers Employee No = 7788.

15) Query to display Name, Salary and Commission for all employees whose Commission Amount is 14 greater than their Salary increased by 5%.

16) Query to display the Current Date.

17) Query to display Name, Hire Date and Salary Review Date which is the 1stMonday after six months of employment.

18) Query to display Name and calculate the number of months between today and the date each employee was hired.

19) Query to display the following for each employee <E-Name> earns <Salary> monthly but wants <3*Current Salary>. Label the Column as Dream Salary.

20) Query to display Name with the 1st letter capitalized and all other letter lower case and length of their name of all the employees whose name starts with ‘J’, ‘A’ and ‘M’.

21) Query to display Name, Hire Date and Day of the week on which the employee started.

22) Query to display Name, Department Name and Department No for all the employees.

23) Query to display Unique Listing of all Jobs that are in Department # 30.

24) Query to display Name, Department Name of all employees who have an ‘A’ in their name.

25) Query to display Name, Job, Department No. and Department Name for all the employees working at the Dallas location.

26) Query to display Name and Employee no. Along with their Manger’s Name and the Manager’s employee no; along with the Employees Name who do not have a Manager.

27) Query to display Name, Department No. And Salary of any employee whose department No. and salary matches both the department no. And the salary of any employee who earns a commission.

28) Query to display Name and Salaries represented by asterisks, where each asterisk (*) signifies $100.

29) Query to display the Highest, Lowest, Sum and Average Salaries of all the employees.

30) Query to display the number of employees performing the same Job type functions.

31) Query to display the no. of managers without listing their names.

32) Query to display the Department Name, Location Name, No. of Employees and the average salary for all employees in that department.

33) Query to display Name and Hire Date for all employees in the same dept. as Blake.

34) Query to display the Employee No. And Name for all employees who earn more than the average salary.

35) Query to display Employee Number and Name for all employees who work in a department with any employee whose name contains a ‘T’.

36) Query to display the names and salaries of all employees who report to King.

37) Query to display the department no, name and job for all employees in the Sale

Create Not only SQL(NoSQL) database (MongoDB) of the above schema and execute the query to obtain the same results.

Semester – II

DS C201	Foundation of Data Science and Analytics	4	20	80	100
DS C202	Cloud Computing	4	20	80	100
DS C203	Business Analytics	4	20	80	100
DS C204	Internet of Things	4	20	80	100
DS P205	Lab II – IoT and Cloud Computing (PRACTICAL)	4	20	80	100
DS VAC206	**Data Analysis with Power BI**	**NIL**	**Non Credit**		**NIL**

SEMESTER - II

COURSE CODE: DS C201	Foundation of Data Science & Analytics
CORE COURSE	4 CREDITS

Unit–1

Definition of Big Data, Big data characteristics & considerations, Data Repositories – analyst perspective, Business drivers for analytics, Typical analytical architecture, Business Intelligence Vs Data Science, Drivers of Big Data analytics, Role of data scientist in Big data ecosystem, Application of Big data analytics.

Unit-2

Need of Data analytic lifecycle, Key roles for successful analytic project, various phases of Data analytic lifecycle: Discovery, Data Preparation, Model Planning, Model Building, Communicating Results, Operationalization.

Unit–3

Overview of Clustering, K- means, Association Rules, Apriori Algorithm, Linear Regression, Logistic Regression.

Unit–4

Naïve Bayesian Classifier, Decision Tress, Time Series analysis, Text Analysis.

TEXT BOOK:

1) David Dietrich, Barry Hiller, “Data Science & Big Data Analytics”, EMC education services, Wiley publications, 2012
2) Trevor Hastie, Robert Tibshirani, Jerome Friedman, "The Elements of Statistical Learning", Springer, Second Edition, 2011.

COURSE CODE: DS C202	Cloud Computing
CORE COURSE	4 CREDITS

Unit–1

Cloud computing definition, Private, public and hybrid cloud, Types of cloud services: IaaS, PaaS, SaaS, Benefits and challenges of cloud computing, Evolution of cloud computing, Usage scenarios and applications, Business models around cloud, Major players in cloud computing, Issues in cloud, Eucalyptus, Nimbus, Open Nebula, CloudSim.

Unit–2

Software as a Service, Platform as a Service, Infrastructure as a Service, Database as a Service, Monitoring as a Service, Communication as a Service, Service providers: Google App Engine, Amazon EC2, Microsoft Azure, Sales force, Introduction to MapReduce, GFS, HDFS, Hadoop Framework.

Unit–3

Collaborating on Calendars, Schedules and Task Management, Collaborating Event Management, Contact Management, Project Management, Collaborating on Word Processing, Database: Storing and Sharing Files, Collaborating via Web-based communication tools, Evaluating Web mail Service, Collaborating via Social Networks, Collaborating via Blogs and Wikis.

Unit–4

Need for Virtualization, Pro and Cons of Virtualization, Types of Virtualization, System VM, Process VM, Virtual Machine Monitor, Virtual Machine Properties, Interpretation and binary Translation, HLL VM, Hypervisors, Xen, KVM, VMWare, Virtual Box, Hyper – V.

Cloud security challenges, Software as a Service Security, Common Standards, The Open Cloud Consortium, The Distributed Management Task Force, Standards for application developers, Standards for Messaging, Standards for Security, End user access to cloud computing, Mobile Internet device and the cloud.

TEXT BOOKS:

1) Cloud Computing for Dummies, by J. Hurwitz, R. Bloor, M. Kanfman, and F. Haiper, Wiley India Edition, 2010 (Unit I).
2) Cloud Computing Implementation Management and Security by J. Ritting house and J. Ransome, CRC Press, 2010 (Unit II).
3) Cloud Computing: A Practical Approach by A. Velte, T. Velte and R. Elsenpeter, Tata McGraw Hill, 2009 (Unit II).
4) Cloud Computing: Web-based Applications That Change the Way You work and Collaborate Online by M. Miller, Que Publishing, August 2008 (Unit III).
5) Virtual Machines by J. E. Smith and R. Nair, Morgan Kaufmann Publishers, 2006 (Unit IV).
6) http://cloud-standards.org/wiki/index.php?title=Main_Page (Unit –V).

REFERENCE BOOKS:

1) Architecting the Cloud: Design Decisions for Cloud Computing Service Models (SaaS, PaaS, and IaaS), by M. Kavis, Wiley, 2014.
2) Mastering In Cloud Computing by R. Buyya, C. Vecchiola and T. Selvi, Tata Mcgraw-Hill Education, 2013.
3) Cloud Computing: SaaS, PaaS, IaaS, Virtualization, Business Models, Mobile, Security and more by K. Jamsa, Jones & Bartlett Learning Company LLC, 2013.

COURSE CODE: DS C203	Business Analytics
CORE COURSE	4 CREDITS

Course Objective:

The objective of the course is to provide an understanding of Basic concepts of Business Analytics like Descriptive, Predictive and Prescriptive Analytics

Learning Outcomes:

1. To learn the complexity of data in business domain.
2. To understand various data modeling and their usage in business.
3. To choose best decision based on various decision support system .techniques.

Module I: Introduction to Business Analytics [6 Hours]

Decision Making Process, Definition of Business Analytics, Categories of Business Analytical Methods and Models, Business Analytics in Practice and Case Studies in - Finance, Human Resource, Marketing, Health Care, Supply Chain, Big Data-Overview of using Data, Types of Data.

Module II: Descriptive Analytics and Data Visualization [8 Hours]

Overview of Description Statistics Central Tendency, Variability, Data Distributions, Association, Data Visualization- Definition, Visualization Techniques –Tables, Cross Tabulations, Charts, Data Dashboards Design.

Module III: Predictive Analytics [10 Hours]

Time Series Analysis and Forecasting Techniques, Data Mining -Definition, Approaches in Data Mining- Data Sampling, Data Preparation, Data Exploration & Reduction, Unsupervised Learning (Classification, Association), Supervised Learning (Data Partitioning, Accuracy, k-Nearest Neighbors, Classification Tree, Regression Tree).

Module IV: Prescriptive Analytics [12 Hours]

Overview of Linear Optimization, Applications of Linear Optimization, Integer Optimization, Decision Analysis.

TEXT BOOK:

1) Camm, Cochran, Fry, Ohlmann, Anderson, Sweeney, Williams-Essentials of Business Analytics, Cengage Learning.

REFERENCE BOOKS:

2) James Evans, Business Analytics, Pearson, Second Edition, 2017.
3) Albright Winston, Business Analytics-Data Analysis-Data Analysis and Decision Making, Cengage Learning, Reprint 2016.
4) Sahil Raj, Business Analytics, Cengage Learning

COURSE CODE: DS C204	Internet of Things
CORE COURSE	4 CREDITS

Unit–1

Introduction to Internet of Things, Definitions and Characteristics of IoT, Physical Design of IoT, Things in IoT, IoT Protocols, Logical Design of IoT, IoT Functional Blocks, IoT Communication Models, IoT Communication APIs, IoT Enabling Technologies - Wireless Sensor Networks, Cloud Computing, Big Data Analytics, Communication Protocols, Embedded Systems.

Textbook 1: 1.1 –1.5

Unit–2

IoTlevels and Development Templates, IoT Level-1, IoT Level-2, IoT Level-3, IoT Level-4, IoT Level-5, IoT Level-6.

IoT and M2M, Introduction, M2M, Difference between IoT and M2M, SDN and NFV for IoT - Software Defined Networking, and Network Function Virtualization, IoT Platform Design Methodology - Introduction, IoT Design Methodology, Step1: Purpose and requirement specification, Step2: Process Specification, Step 3: Domain Model Specification, Step 4: Information Model Specification, Step 5: Service Specification, Step 6: IoT Level Specification, Step 7: Function View Specification, Step 8: Operational View Specification, Step 9: Device and Component Integration, Step 10: Application Development, IoT System Logical Design Using Python.

Textbooks 1: 3.1-3.4, 5.1-5.4, 6.1-6.11

Unit-3

IoT Physical Devices and End Points: What is an IoT Device, Exemplary Device Raspberry Pi, About the Board, Linux on Raspberry Pi, Raspberry pi interfaces, programming raspberry pi with python, IoT physical servers and cloud offerings - introduction to cloud storage models and communication Networks, wamp-autobahn for IoT, xively cloud for IoT, python web application frame work-django, designing a RESTful web API.

Textbook 1: 7.1-7.7, 8.1-8.7

Unit–4

Data Analytics for IoT; Introduction AppacheHadoop, using HadoopMapReduce for Batch Data Analysis.

Textbook 1: 10.1 -10.8

Ethics: Characterizing the IoT, Privacy, Control, Distributing Control and Crowd Sourcing, Environment, Physical Thing, Electronics, InternetService, Solutions, Internet of Things as Part of Solution, Cautious Optimizing, The Open IoT definition.

Textbook 2: Chapter 11

TEXT BOOKS:

1) Internet Of Things-A Hands on Approach, by Arshdeep Bahga and Vijay Madisetti, University of Penn, http://www.internet-of-things-book.com/
2) Designing the Internet of Things, by Adrian McEwen and Hakim Cassimally, Wiley Publication.

REFERENCES:

1) Internet of Things: Converging Technologies for Smart Environments and Integrated Ecosystems. By Ovidiu Vermesan and Peter Friess, River Publishers Series in Communication.

DS P205 Lab. II: IoT and Cloud Computing (4 Credits)

IoT : List of Experiments

1. Define and Explain Eclipse IoT Project.
2. List and summarize few Eclipse IoT Projects.
3. Sketch the architecture of IoT Toolkit and explain each entity in brief.
4. Demonstrate a smart object API gateway service reference implementation in IoT toolkit.
5. Write and explain working of an HTTP- to-CoAP semantic mapping proxy in IoT toolkit.
6. Describe gateway-as-a-service deployment in IoT toolkit.
7. Explain application framework and embedded software agents for IoT toolkit.
8. Explain working of Raspberry Pi.
9. Connect Raspberry Pi with your existing system components.
10. Give overview of Zetta.

Design based Problems (DP)/Open Ended Problem:

1. How do you connect and display your Raspberry Pi on a Monitor Or TV?
2. Create any circuitry project using Arduino.

Major Equipment:

Raspberry pi, Arduino

List of Open Source Software/learning website:

1. https://github.com/connectIOT/iottoolkit
2. https://www.arduino.cc/
3. http://www.zettajs.org/
4. Contiki (Open source IoT operating system)
5. Arduino (open source IoT project)
6. IoT Toolkit (smart object API gateway service reference implementation)
7. Zetta (Based on Node.js, Zetta can create IoT servers that link to various devices and sensors)

Cloud Computing : List of Experiments

1. Create virtual machines that access different programs on same platform.
2. Create virtual machines that access different programs on different platforms.
3. Exploring Google cloud for the following
   a. Storage
   b. Sharing of data
   c. Manage your calendar, to-do lists
   d. A document editing tool
4. Exploring Microsoft cloud
5. Exploring Amazon cloud
6. Exploring Open cloud

Course Code: DS VAC206	Data Analysis with Power BI

Course Description:

This course is designed to provide an introduction to the Power BI tool for data analysis and visualization. Students will learn how to import, clean, and transform data, create relationships between tables, and create visualizations and dashboards. Additionally, they will learn how to share and collaborate on their work using Power BI.

Prerequisites:

- Basic understanding of data analysis and visualization
- Familiarity with Excel or another spreadsheet application

Course Objectives:

1. Understanding the basics of data analysis and visualization with Power BI.
2. Familiarize with the Power BI user interface and data model components such as tables, relationships, and measures.
3. Learning how to import, clean, and transform data from different sources using Power Query.
4. Learning how to create effective and interactive data visualizations such as charts, graphs, and maps with Power BI.

Course Outline:

Module - I

Introduction to Power BI, Overview of Power BI, Installing Power BI, importing data into Power BI, Data Preparation in Power BI, Data cleaning and transformation, creating relationships between tables, Data modeling best practices, Creating Visualizations in Power BI.

Module - II

Overview of visualizations in Power BI, Creating and formatting charts, tables, and matrices, Creating maps, gauges, and cards, Advanced Visualizations in Power BI, Creating custom visuals, Using Power BI visuals from the marketplace, Creating and using drill-throughs and hierarchies.

Module - III

Creating dashboards in Power BI, Designing and building dashboards, Adding filters and slicers to dashboards, Configuring dashboard options, Sharing and Collaborating with Power BI, Sharing dashboards and reports, Using Power BI workspaces, Collaborating with others in Power BI, Data Analysis with Power BI.

Module - IV

Case Study

Applying Power BI to a real-world data analysis problem, importing data and performing data cleaning, creating visualizations and dashboards, and Sharing and collaborating on the project.

Reference Books:

1) "Beginning Power BI: A Practical Guide to Self-Service Data Analytics with Excel 2016 and Power BI Desktop" by Dan Clark.
2) "Mastering Microsoft Power BI: Expert Techniques for Effective Data Analytics and Business Intelligence" by Brett Powell.
3) "Applied Microsoft Power BI (2nd Edition): Bring your data to life!" by Teo Lachev.

Semester – III

DS C301	Machine Learning	4	20	80	100
DS C302	Big Data Analytics	4	20	80	100
DS CBCS *	Data Analytics	4	20	80	100
ELECTIVES					
DS E303	Predictive Analytics	4	20	80	100
DS E304	Decision Management Systems	4	20	80	100
DS E305	Text Analytics	4	20	80	100
DS P306	Lab III – Python Programming and Machine Learning (PRACTICAL)	4	20	80	100
DS VAC 307	**Programming with NumPy and Pandas**	**NIL**	**Non Credit**		**NIL**

SEMESTER - III

COURSE CODE: DS C301	Machine Learning
CORE COURSE	4 CREDITS

Unit–1

Introduction – Types of Machine Learning, Designing a Learning System, Issues in Machine Learning; The Concept Learning Task - General-to-specific ordering of hypotheses, Find-S, List then eliminate algorithm, Candidate elimination algorithm, Inductive bias - Decision Tree Learning - Decision tree learning algorithm, Instance based Learning, Nearest neighbors method.

Unit–2

Artificial Neural Networks – Perceptrons, Learning rules, Gradient descent and the Delta rule, Adaline, Madaline Network, Multilayer networks, Derivation of Backpropagation rule- Backpropagation Algorithm- Convergence, Generalization; – Evaluating Hypotheses – Estimating Hypotheses Accuracy, Basics of sampling Theory, Radial basis function networks, Support Vector Machine.

Unit-3

Supervised Learning- Linear Regression (Gradient Descent, Normal Equations), Weighted Linear Regression (LWR), Logistic Regression, Generative Models (Gaussian Discriminant Analysis, Naive Bayes), Learning – Bayes theorem, Concept learning, Bayes Optimal Classifier, Naïve Bayes classifier, Bayesian belief networks, Tree Ensembles (Decision trees, Random Forests, Boosting and Gradient Boosting).

Unit–4

Unsupervised Learning- K-means, Gaussian Mixture Model (GMM), Expectation Maximization (EM), Variational Auto-encoder (VAE), Factor Analysis, Principal Components Analysis (PCA), Independent Components Analysis (ICA), Linear Discriminant Analysis (LDA), Vector Quantization – Self Organizing Feature Map.

Reinforcement learning: Markov decision process (MDP), Hidden Markov Model(HMM), Bellman equations, Value iteration and policy iteration, Linear quadratic regulation, Linear Quadratic Gaussian, Q-learning, Monte Carlo Methods.

TEXT BOOKS:

1) T. Mitchell, “Machine Learning” Mcgraw Hill Publisher.
2) T. Hastie, R. Tibshirani, J. Friedman “The Element of Statistical Learning” 2e 2008
3) E. Alpaydin, Introduction to Machine Learning. Eastern Economy Edition, Prentice Hall of India.
4) C. M. Bishop, Pattern recognition and Machine Learning, Springer

COURSE CODE: DS C302	Big Data Analytics
CORE COURSE	4 CREDITS

Unit–1

Linear Methods for Regression and Classification: Overview of supervised learning, Linear regression models and least squares, Multiple regression, Multiple outputs, Subset selection, Ridge regression, Lasso regression, Linear Discriminant Analysis, Logistic regression, Perceptron learning algorithm.

Unit–2

Model Assessment and Selection: Bias, Variance, and model complexity, Bias-variance trade off, Optimism of the training error rate, Estimate of In-sample prediction error, Effective number of parameters, Bayesian approach and BIC, Cross- validation, Boot strap methods, conditional or expected test error.

Additive Models, Trees, and Boosting: Generalized additive models**,** Regression and classification trees, Boosting methods-exponential loss and AdaBoost, Numerical Optimization via gradient boosting, Examples (Spam data, California housing, New Zealand fish, Demographic data).

Unit-3

Neural Networks (NN), Support Vector Machines (SVM), and K-nearest Neighbour: Fitting neural networks, Back propagation, Issues in training NN, SVM for classification, Reproducing Kernels, SVM for regression, K-nearest –Neighbour classifiers(Image Scene Classification)

Unsupervised Learning and Random forests: Association rules, Cluster analysis, Principal Components, Random forests and analysis.

Unit–4

Inferential Statistics and Prescriptive Analytics: Assessing Performance of a classification Algorithm (t-test, McNemar’s test, Paired t-test, paired f-test), Analysis of Variance, Creating data for analytics through designed experiments.

TEXT BOOKS:

1) 1.Trevor Hastie, Robert Tibshirani, Jerome Friedman*, The Elements of Statistical Learning-Data Mining, Inference, and Prediction*, Second Edition , Springer Verlag, 2009.

   [Chapters: 2, 3(3.1-3.4, 3.6), 4 (4.3-4.5), 7 (excluding 7.8 and 7.9), 9 (9.1, 9.2), 10 (10.1-10.5, 10.8, 10.10, 10.14), 11 (11.3-11.6), 12(12.1-12.3), 13.3, 14(14.1-14.3.8, 14.5.1), 15]

2) E.Alpaydin*, Introduction to Machine Learning*, Prentice Hall Of India, 2010,(Chapter-19).

3) G.James, D.Witten, T.Hastie, R.Tibshirani-*An introduction to statistical learning with applications in R*, Springer, 2013. (2.3, 3.6.1-3.6.3, 4.6.1-4.6.3, 5.3, 6.6.1, 8.3.1, 8.3.2, 10.4, 10.5.1)

REFERENCES:

1) C.M. Bishop –Pattern Recognition and Machine Learning, Springer, 2006.
2) L. Wasserman-All of statistics.

COURSE CODE: DS CBCS	Data Analytics
CORE COURSE	4 CREDITS

Module-1:

Introduction to Data Analytics, Relation to data mining, machine learning, Big data and statistics, Business Intelligence (BI) vs. Data Science. Types of Data: Structured v/s unstructured data, Examples of data pre-processing, Quantitative vs qualitative data, Four levels of data. Stages of a data science project: Defining the goal, Data collection and mgmt, Explore the data, Modeling, Model evaluation and critique, Presentation and documentation.

Module-2:

Data Analytic life-cycle, introduction to Linear algebra for data science: Vectors and matrices. Introduction to Probability: Bayesian versus Frequentist, Frequentist approach, the law of large numbers, Random variables, Sampling data, Probability sampling, Random sampling, Unequal probability sampling, measurement of statistics, Measures of center (Mean, Median, Mode, Skewness, Quantile, Percentile), Measures of variation, Measures of relative standing, Correlations in data.

Module - 3:

Basic Algorithms like Linear Regression, k-Nearest Neighbors (k-NN), k-Means. Feature Extraction, Eigen vectors and Eigen values, Principal Component Analysis. Linear and Logistic Regressions.

Module - 4:

Conditional prob, Bayesian ideas revisited, Bayes theorem, More applications of Bayes theorem, Decision Tree and Random Forest. Time series data analysis and Text analysis

TEXT BOOK:

1) David Dietrich, Barry Hiller, “Data Science & Big Data Analytics”, EMC education services, Wiley publications, 2012.
2) Trevor Hastie, Robert Tibshirani, Jerome Friedman, "The Elements of Statistical Learning", Springer, Second Edition, 2011.

REFERENCE BOOKS:

1) Principles of Data Science, Sinan Ozdemir, Packt Publishing Ltd 2016.
2) An Introduction to Statistical Learning with Applications in R. James G, Witten D, Hastie Tibshirani R, Springer, 2013.
3) Hands-On Data Science with R: Techniques to perform data manipulation and ..., Vitor Bianchi Lanzetta, Nataraj Dasgupta, Ricardo Anjoleto Farias, Packt publishing ltd, 2018.

COURSE CODE: DS E303	Predictive Analytics
ELECTIVE COURSE - I	4 CREDITS

Unit–1

Introduction: Prediction versus Interpretation, Key Ingredients of Predictive Models, Predictive Modeling Process.

Data Pre-processing: Data Transformations for Individual Predictors- Centering and Scaling, Transformations to Resolve Skewness, Data Transformations for Multiple Predictors- Transformations to Resolve Outliers, Data Reduction and Feature Extraction, Removing Predictors- Predictor Correlations, Adding Predictors, Binning Predictors.

Unit–2

Over-Fitting and Model Tuning: The Problem of Over-Fitting Model Tuning, Data Splitting, Resampling Techniques. Regression Models- Quantitative Measures of Performance, Linear Regression- Partial Least Squares, Penalized Models, Nonlinear Regression Models - Neural Networks, K-Nearest Neighbors.

Classification Models: introduction of Classification Models - Discriminant Analysis and Other Linear Classification Models - Nonlinear Classification Model -Na¨ıve Bayes - Support Vector Machines . Classification Trees and Rule-Based Models- Basic Classification Trees , Rule-Based Models -PART, Bagged Trees, Boosting- Ada Boost

Unit–3

Introduction to Feature Selection -Consequences of Using Non-informative Predictors- Approaches for Reducing the Number of Predictors-Factors That Can Affect Model Performance- Measurement Error in the Outcome, Measurement Error in the Predictors.

Unit–4

Predicting Cognitive Impairment Predicting Caravan Policy Ownership, The Effect of Class Imbalance- Sampling Methods-Cost-Sensitive Training-Job Scheduling

TEXT BOOKS:

1) Kuhn, Max, Kjell Johnson. “*Applied predictive modeling*”. Springer, 2018**.**

REFERENCES:

1) Siegel, Eric. “*Predictive analytics”: The power to predict who will click, buy, lie, or die*. John Wiley & Sons, 2013.
2) Abbott, Dean. “*Applied predictive analytics”: Principles and techniques for the professional data analyst*. John Wiley & Sons, 2014.
3) Miner, Gary, “Practical text mining and statistical analysis for non-structured text data applications”. Academic Press, 2012.

COURSE CODE: DS E304	Decision Management Systems
ELECTIVE COURSE - I	4 CREDITS

Unit–1

Introduction Artificial Intelligence, Digital Decisioning, Digital Decisioning Principles.

Unit–2

Discover and Model Decisions - Characteristics of Suitable Decisions - A Decision Taxonomy - Finding Decisions - Documenting Decisions - Prioritizing Decisions.

Design and Implement Decision Services - Build Decision Services - Integrate Decision Services - Best Practices for Decision Services Construction.

Unit–3

Monitor and Improve Decisions - What Is Decision Analysis? - Monitor Decisions - Determine the Appropriate Response - Develop New Decision-Making Approaches - Confirm the Impact Is as Expected - Deploy the Change.

Unit–4

Enablers for Decision Management Systems: People Enablers, Process Enablers, Technology Enablers.

TEXT BOOKS:

1) James Taylor, “Decision Management Systems-A Practical guide to using Business rules and Predictive Analytics”, IBM Press, 2012.
2) James Tayler, ``Digital decisioning: Using Decisioning Management to deliver Business Impact on IA, Meghan-Kiffer Press, 2019.

COURSE CODE: DS E305	Text Analytics
ELECTIVE COURSE - I	4 CREDITS

Unit–1

Text Mining - Definition-General Architecture–Algorithms– Core Operations : Distributions-Frequent and near Frequent Sets- Associations – Isolation of Interesting Patterns – Analyzing Document Collection over Time – Using Background Knowledge for text mining–Pre-processing-Textual information to numerical vectors -Collecting documents-document standardization-tokenization-lemmatization-vector generation for prediction-sentence boundary determination -evaluation performance

Unit–2

Text Categorization –Definition –Document Representation – Feature Selection -Decision TreeClassifiers -Rule- based Classifiers - Probabilistic and Naive Bayes Classifiers – Linear Classifiers-Classification of Linked and Web Data - Meta-Algorithms–Clustering –Definition-Vector Space Models - Distance-based Algorithms-Word and Phrase-based Clustering -Semi-Supervised Clustering -Transfer Learning

Unit–3

Information retrieval and text mining-keyword search-nearest -neighbor methods-similarity-web-based document search-matching-inverted lists-evaluation. Information extraction-Architecture -Co-reference –Named Entity and Relation

Extraction-Templatefilling and database construction– Applications. Inductive -Unsupervised Algorithms for Information Extraction. Text Summarization Techniques -Topic Representation -Influence of Context -Indicator representations -Pattern Extraction -Apriori Algorithm – FP Tree algorithm

Unit–4

Probabilistic Models for Text Mining -Mixture Models -Stochastic Processes in Bayesian Nonparametric Models -Graphical Models -Relationship Between Clustering, Dimension Reduction and Topic Modeling -Latent Semantic Indexing -Probabilistic Latent Semantic Indexing -Latent Dirichlet Allocation-Interpretation and Evaluation -Probabilistic Document M.Sc Data Clustering and Topic Models -Probabilistic Models for Information Extraction -Hidden Markov Models -Stochastic Context-Free Grammars - Maximal Entropy Modeling -Maximal Entropy Markov Models -Conditional Random Fields

TEXT BOOKS:

1) Sholom Weiss, Nitin Indurkhya, Tong Zhang, Fred Damerau “The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data”, Springer, paperback 2010

REFERENCES:

1) Ronen Feldman, James Sanger-“The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data”-Cambridge University press, 2006.
2) Charu C. Aggarwal, Cheng Xiang Zhai, Mining Text Data, Springer; 2012.

Ds P306 Lab. III: Machine Learning and Python Programming

Machine Learning: List of Experiments

1. Basic operations in Python implementation.
2. Loading data from Training set and testing the Models.
3. Learn to predict values with Linear Regression.
4. Learn to predict states using Logistic Regression.
5. Learn the definition of a Perceptron as a building block for neural networks, and the perceptron algorithm for classification.
6. Learn the definition of a Neural Network, Learn to train them using Backpropagation network.
7. Train Decision Trees to predict states and classification.
8. Learn the Bayes rule, and how to apply it to predicting data using the Naive Bayes algorithm.
9. Learn to train a Support Vector Machine to separate data linearly.
10. Use Kernel Methods in order to train SVMs on data that is not linearly separable.
11. Learn the basics of clustering Data, Cluster data with the K-means algorithm.
12. Cluster data with Gaussian Mixture Models.
13. Optimize Gaussian Mixture Models with Expectation Maximization.
14. Learn to scale features in your data, Learn to select the best features for training data.
15. Reduce the dimensionality of the data using Principal Component Analysis and Independent Component Analysis and LDA
16. Learn how to define Markov Decision Processes to solve real-world problems.
17. Learn about policies and value functions, Derive the Bellman Equations.
18. Write your own implementations of iterative policy evaluation, policy improvement, policy Iteration, and value Iteration.
19. Implement classic Monte Carlo prediction and control methods.
20. Learn how to tune hyper parameters of an estimator.
21. Plotting of Validation curve and learning curve to evaluate the model.
22. Evaluating Estimator performance, Cross validation

Python Programming

Features of Python, Installing Python for windows and setting up paths, writing and Executing of a python programs, Python Virtual machine, Frozen binaries, Comparison between C, Java and python, Comments, Docstrings, How python sees variables, Data types in Python, built in types, sequences in python, sets, literals in Python, user defined data types, identifiers & reserved words, Naming convention in python,

Various Operators in Python, Input & Output, Control statements, if statements, while loop, for loop, infinite loop, nested loop, else suit, break, continue, pass ,assert, return statements, command line arguments.

Arrays in python, advantages using arrays, creating arrays, importing the array module, indexing and slicing on arrays, Processing the arrays, Comparing arrays.

Strings in Python, Creating strings, Length of a string, Indexing in strings, Slicing strings, Concatenation and Comparing strings, Finding SubStrings, Replacing a String.

Functions in Python, Define a function, Calling a function, return from function, pass by object Reference, Positional arguments, Default arguments, Recursive functions.

Inheritance: Define inheritance, types of inheritance, constructors in inheritance, overriding super class constructors & methods, the super() method, MRO

Polymorphism: Duck typing philosophy of Python, operator overloading, method overriding, interfaces in python

Exceptions: Errors in a python program, Exceptions, Exception handling, Types of Exceptions, The Exception block, the assert statement, user defined exceptions

Python Programming : List of Experiments

1. Write a menu driven program to convert the given temperature from Fahrenheit to Celsius and vice versa depending upon users choice.
2. Write a Program to calculate total marks, percentage and grade of a student. Marks obtained in each of the three subjects are to be input by the user. Assign grades according to the following criteria:

   Grade A: Percentage >=80

   Grade B: Percentage>=70 and <80

   Grade C: Percentage>=60 and <70

   Grade D: Percentage>=40 and <60

   Grade E: Percentage<40
3. Write a menu-driven program, using user-defined functions to find the area of rectangle, square, circle and triangle by accepting suitable input parameters from user.
4. Write a Program to display the first n terms of Fibonacci series.
5. Write a Program to find factorial of the given number.
6. Write a Program to find sum of the following series for n terms: 1 – 2/2! + 3/3! - - - - - n/n!
7. Write a Program to calculate the sum and product of two compatible matrices.
8. Install MySQL and connector. Write Python programs to retrie, inserting, delete, update rows in a table.

Course Code: DS VAC307	Programming with NumPy and Pandas

Course Description:

This course is designed to provide an in-depth understanding of the NumPy and Pandas libraries for data analysis in Python. Students will learn how to use NumPy and Pandas to manipulate and analyze data, including reading data from various file formats, selecting and filtering data, and performing calculations and aggregations.

Prerequisites:

- Proficiency in Python programming language
- Basic knowledge of data types such as lists, dictionaries, and arrays
- Basic understanding of data analysis and manipulation

Course Objectives:

1) Understanding the basics of data manipulation and analysis with Python.
2) Familiarize with the fundamentals of NumPy and Pandas libraries, and how to use them for data manipulation, analysis, and visualization.
3) Understanding and implementing basic and advanced data wrangling operations with NumPy and Pandas.
4) Implementing statistical analysis of data with NumPy and Pandas.

Course Outline:

Module - I

Introduction to NumPy and Pandas, Overview of NumPy and Pandas, Installing NumPy and Pandas, Importing NumPy and Pandas and basic commands, NumPy Array Operations**,** Creating and manipulating NumPy arrays, Indexing and selecting data, Broadcasting and vectorization, Linear algebra operations, Pandas Data Structures

Module - II

Series and Data Frames, Indexing and selecting data, filtering data, Merging and joining data, Data Input and Output, Reading and writing data in various file formats (CSV, Excel, JSON), Data Cleaning and Manipulation with Pandas. Handling missing data, handling duplicates, changing data types, renaming columns, Adding and deleting columns, and Grouping data.

Module - III

Aggregating data, and Advanced NumPy and Pandas Operations. Reshaping and pivoting data, applying functions to data, Handling time-series data, Advanced indexing and selection, Handling multi-index data, and Data Visualization.

Module - IV

Introduction to data visualization using Pandas and Matplotlib, plotting data using line, bar, and scatter plots Creating histograms and box plots.

Reference Books:

1) "Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython" by Wes McKinney.
2) "Python Data Science Handbook: Essential Tools for Working with Data" by Jake VanderPlas.
3) "Data Wrangling with Pandas: Tips and Tools to Make Your Life Easier" by Kevin Markham.

Semester – IV

DS C401	Social Media Analytics	4	20	80	100
DS C402	Time Series Analysis and Forecasting	4	20	80	100
ELECTIVES					
DS E403	Healthcare Analytics	4	20	80	100
DS E404	Sentiment Analytics	4	20	80	100
DS E405	Image and Video Analytics	4	20	80	100
DS PR406	Project Work / Dissertation	8			200

SEMESTER - IV

COURSE CODE: DS C401	Social Media Analytics
CORE COURSE	4 CREDITS

Unit–1

Introduction to Social Network Analysis, Mathematical representations of Social Networks: Notations for Social Network data – Graph theoretic, sociometric. Graphs – Subgraphs, Dyads, Triads, Nodal degree, Density, Walks, trails and paths, Connected graphs and components, Geodesics, distance and diameter, Connectivity, Isomorphic graphs and subgraphs.

Unit–2

Directed graphs – Dyads, Nodal indegree and outdegree, Density, directed walks, paths and semi paths, Reachability and connectivity, Geodesics, distance and diameter. Signed graphs and signed directed graphs Matrices – for graphs, digraphs, valued graphs, two-mode networks, Basic matrix operations, Computing simple network properties

Unit–3

Centrality: Actor centrality, Nondirectional relationships – degree, closeness, betweenness centrality, Directional relations – centrality

Structural relationships – strong and weak ties, homophily, positive and negative relationships, Link analysis.

Unit–4

Network dynamics – cascading behavior, small-world phenomenon, epidemics. Tools for Social Network Analysis - UCINET-PAJEK-ETDRAW-StOCNET- Splus-R-NodeXL-SIENA and RSIENA-Real world Social Networks (Facebook-Twitter etc.)

TEXT BOOKS:

1) Social Network Analysis: Methods and Applications, Book by Katherine Faust and Stanley Wasserman.
2) Networks, Crowds, and Markets: Reasoning about a Highly Connected World Book by David Easley and Jon Kleinberg.
3) Social and Economic Networks Book by Matthew O. Jackson.

COURSE CODE: DS C402	**Time Series Analysis and Forecasting**
CORE COURSE	**4 CREDITS**

Unit–1

Examples of Nature of Time series data – Time series statistical models – Measures of dependence -Stationary. Time series regression – Detrending and differencing – Smoothing a time series

Unit–2

Auto Regressive models – Moving Average models - ARMA models – Auto Correlation Function - Partial Auto Correlation Function – Forecasting algorithms – Estimation: Yule-Walker, Method of moments, MLE and LSE

Basics of ARIMA models: andom models with drift, Steps to fitting ARMA model – Multiplicative Seasonal ARIMA models: Mixed, SARMA – Generalized Auto Regressive Conditionally Heteroscedastic (GARCH) models

Unit–3

Cyclical Behaviour and Periodicity: Concepts, Periodic Series, Star Magnitude - The Spectral Density: Periodic stationary process–Periodogram: Spectral analysis as ANOVA, Principal Component Analysis

Unit–4

Dynamic Linear Models – Examples of DLMs – Filtering DLM – Smoothing DLM: Kalman, Lag One covariance – Forecasting DLM – Maximum Likelihood Estimator for DLMs

TEXT BOOKS:

1) Shumway and Stoffer. Time Series Analysis and its applications, with examples in R. 4ed, Springer. 2016.

REFERENCES

1) Brockwell & Davis. Introduction to Time Series and Forecasting, 3rd edition, Springer. 2016.
2) Cryer & Chan. Time Series Analysis with Applications in R, Springer. 2008
3) Prado & West. Time Series: Modeling, Computation, and Inference Chapman & Hall. 2010.
4) Petris, Petrone, Campagnoli. Dynamic Linear Models with R, Springer. 2009
5) Ruppert & Matteson. Statistics and Data Analysis for Financial Engineering with R examples, 2ed, Springer.

COURSE CODE: DS E403	Healthcare Analytics
ELECTIVE COURSE - II	4 CREDITS

Unit-I

Introduction: Introduction to Healthcare Data Analytics- Electronic Health Records– Components of EHR- Coding Systems- Benefits of EHR- Barrier to Adopting EHR- Challenges- Phenotyping Algorithms.

Analysis: Biomedical Image Analysis- Mining of Sensor Data in Healthcare- Biomedical Signal Analysis- Genomic Data Analysis for Personalized Medicine.

Unit-II

Analytics: Natural Language Processing and Data Mining for Clinical Text- Mining the Biomedical-Social Media Analytics for Healthcare.

Unit-III

Advanced Data Analytics: Advanced Data Analytics for Healthcare– Review of Clinical Prediction Models- Temporal Data Mining for Healthcare Data- Visual Analytics for Healthcare- Predictive Models for Integrating Clinical and Genomic Data- Information Retrieval for Healthcare- Privacy-Preserving Data Publishing Methods in Healthcare.

Unit-IV

Applications: Applications and Practical Systems for Healthcare– Data Analytics for Pervasive Health-Fraud Detection in Healthcare- Data Analytics for Pharmaceutical Discoveries- Clinical Decision Support Systems- Computer-Assisted Medical Image Analysis Systems- Mobile Imaging and Analytics for Biomedical Data.

TEXT BOOKS:

1) Chandan K. Reddy and Charu C Aggarwal, “Healthcare data analytics”, Taylor & Francis, 2015

REFERENCE:

1) Hui Yang and Eva K. Lee, “Healthcare Analytics: From Data to Knowledge to Healthcare Improvement, Wiley, 2016.

COURSE CODE: DS E404	**Sentiment Analytics**
ELECTIVE COURSE - II	**4 CREDITS**

Unit–1

Sentiment Analysis Applications: Sentiment Analysis Research - Opinion Spam Detection. The Problem of Sentiment Analysis: Problem Definitions - Opinion Summarization – Different Types of Opinions - Subjectivity and Emotion - Author and Reader Standing Point. Document Sentiment Classification: Sentiment Classification Using Supervised Learning – Sentiment Classification Using Unsupervised Learning - Sentiment Rating Prediction – Cross-Domain Sentiment Classification -Cross-Language Sentiment Classification.

Unit–2

Sentence Subjectivity and Sentiment Classification: Subjectivity Classification – Sentence Sentiment Classification - Dealing with Conditional Sentences - Dealing with Sarcastic Sentences – Cross language Subjectivity and Sentiment Classification - Using Discourse Information for Sentiment Classification.

Unit–3

Aspect-based Sentiment Analysis: Aspect Sentiment Classification - Basic Rules of Opinions and Compositional Semantics - Aspect Extraction - Identifying Resource Usage Aspect -Simultaneous Opinion Lexicon Expansion and Aspect. Extraction: Grouping Aspects into Categories - Entity, Opinion Holder and Time Extraction - Co reference Resolution and Word Sense Disambiguation.

Unit–4

Sentiment Lexicon Generation: Dictionary-based Approach - Corpus-based Approach - Desirable and Undesirable Facts. Opinion Summarization: Aspect-based Opinion Summarization - Improvements to Aspect-based Opinion Summarization - Contrastive View Summarization - Traditional Summarization Analysis of Comparative Opinions: Problem Definitions - Identify Comparative Sentences - Identifying Preferred Entities.

Opinion Search and Retrieval: Web Search vs. Opinion Search - Existing Opinion Retrieval Techniques Opinion Spam Detection: Types of Spam and Spamming - Supervised Spam Detection -Unsupervised Spam Detection - Group Spam Detection.

TEXT BOOKS:

1) Sentiment Analysis and Opinion Mining (Synthesis Lectures on Human Language Technologies), Bing Liu, Morgan & Claypool Publishers (2012)

REFERENCES:

1) Sentiment Analysis: Mining Opinions, Sentiments, and Emotions, Bing Liu, Cambridge University Press (2015)
2) http://nptel.ac.in/courses/106105158/61
3) Sentiment Analysis: Second Edition, Gerardus Blokdyk, Createspace Independent Publishing Platform (2018)

COURSE CODE: DS E405	Image and Video Analytics
ELECTIVE COURSE - II	4 CREDITS

Unit–1

Introduction to Big Data Platform – Challenges of Conventional Systems – Web Data – Evolution of Analytic Scalability – Analytic Processes and Tools – Analysis Vs Reporting – Modern Data Analytic Tools – Data Analysis: Regression Modeling – Bayesian Modeling – Rule Induction.

Unit–2

Introduction to Stream Concepts – Stream Data Model And Architecture – Stream Computing – Sampling Data in a Stream – Filtering Streams – Counting Distinct Elements in a Stream– Estimating Moments – Counting Oneness in a Window – Decaying Window – Real Time Analytics Platform(RTAP) Applications – Case Studies

Unit–3

Introduction – Video Basics – Fundamentals for Video Surveillance – Scene Artifacts – Object Detection and Tracking: Adaptive Background Modelling and Subtraction – Pedestrian Detection and Tracking – Vehicle Detection and Tracking – Articulated Human Motion Tracking in Low Dimensional Latent Spaces.

Event Modelling – Behavioural Analysis – Human Activity Recognition – Complex Activity Recognition – Activity modeling using 3D shape - Video summarization – shape based activity models – Suspicious Activity Detection**.**

Unit–4

Introduction: Overview of Recognition algorithms – Human Recognition using Face: - Face Recognition from still images – Face Recognition from video – Evaluation of Face Recognition Technologies – Human Recognition using gait: HMM Framework for Gait Recognition – View Invariant Gait Recognition – Role of Shape and Dynamics in Gait Recognition

REFERENCES:

1) Michael Berthold, David J.Hand, “Intelligent Data Analysis”, Springer, 2007.
2) Anand Rajaraman and Jeffrey David Ullman, “Mining of Massive Datasets”, Cambridge University Press, 2012.
3) Yunqian Ma, Gang Qian, “Intelligent Video Surveillance: Systems and Technology”, CRC Press (Taylor and Francis Group), 2009.
4) Rama Chellappa, Amit K.Roy– Chowdhury, Kevin Zhou.S, “Recognition of Humans and their Activities using Video”, Morgan & Claypool Publis1.hers, 2005.

COURSE CODE: DS PR406	**PROJECT / DISSERTATION (8 CREDITS)**