MASTER OF SCIENCE
IN
COMPUTER SCIENCE

(DATA ANALYTICS)

PROGRAMME STRUCTURE AND SYLLABUS

From 2020-21 Admission onwards

BOARD OF STUDIES IN COMPUTER APPLICATIONS (PG)

MAHATMA GANDHI UNIVERSITY

KOTTAYAM

BOARD OF STUDIES IN COMPUTER APPLICATIONS (PG)

1	Krishnakumar M R Associate Professor and HOD Department of Computer Application SAS SNDP Yogam College, Konni	Chairman
2	Dr. Manu Sankar Assistant Professor Department of Computer Applications Sree Sankara Vidhyapeetom College, Valayanchirangara	Member
3	Sreekumar B Associate Professor , Department of Computer Applications, NSS College Rajakumari	Member
4	Dr. Sabu M K Professor Department of Computer Applications, CUSAT	Member
5	Biju Skaria Associate Professor, Department of Computer Applications Mar Athanasius College of Engineering, Kothamangalam	Member
6	Amruth K John Associate Professor Department of Computer Applications Marian College Kuttikkanam (Autonomous)	Member
7	Spasiba Raveendran Assistant Professor Department of Computer Science SAS SNDP College, Konni	Member
8	Jasir M P Assistant Professor Department of Computer Applications MES College, Marampally, Alua	Member
9	Dr Benymol Jose Associate Professor Department of Computer Applications Marian College Kuttikkanam, Idukki	Member
10	Shyni S Das Associate Professor Department of Computer Applications SAS SNDP Yogam College, Konni	Member
11	Maya N Associate Professor Department of Computer Applications NSS Hindu College, Changanacherry	Member

1. **Aim of the Programme**

The Master's programme in Computer Science with specialization in Data Analytics aims to combine a scientific mind set with specialist technical knowledge, enabling graduates to analyse, design, validate and implement state-of-the-art ICT systems in their operational context. It is a broad-based program that covers concepts from engineering, science and business with the aim of producing high-quality software professionals.

2. **Eligibility For Admission**

The eligibility for admission to M.Sc. Computer Science with Data Analytics programme in affiliated institutions under Mahatma Gandhi University is a regular B.Sc. Degree with Mathematics /Computer Science /Electronics as one of the subjects (Main or Subsidiary) or BCA/B.Tech degree with not less than 50% marks.

Note: Candidates having degree in Computer Science/ Computer Application/ IT/Electronics shall be given a weightage of 20% in their qualifying degree examination marks considered for ranking for admission to M.Sc. Computer Science with Data Analytics.

3. **Programme Structure and Duration**

The duration of the programme shall be 4 semesters. The duration of each semester shall be 90 working days. Odd semesters from June to October and even semesters from December to April.

4. **Examination**

There shall be University examination for theory and practical at the end of each semester. Main Project evaluation and Comprehensive Viva -Voce shall be conducted at the end of the programme only. Comprehensive viva-voce in the fourth semester will cover entire courses in the programme. Project evaluation and Viva-Voce shall be conducted by two external examiners and one internal examiner. Mini project evaluation of second and third semester is done along with university practical examination. The same is conducted by external examiner appointed from university. End-semester examination of all courses except project will be of three hours duration.

5. **Faculty under which degree is awarded**

Faculty of Science

6.   Curriculum Design Abstract

Semester I

CA030101	- Statistics for Data Analytics
CA030102	- Introduction to Data Analytics and Machine Learning
CA030103	- Advanced Operating Systems
CA030104	- Data Structure using C
CA030105	- Python Programming for Analytics
CA030106	- Python and Data structure Lab

Semester II

CA030201	- Mathematics for Data Analytics
CA030202	- Advanced Database Management system
CA030203	- Data Mining and Analytics
CA030204	- Programming with Java
CA030205	- Java & SQL Lab
CA030206	- Mini Project I

Semester III

CA030301	- Statistical Modeling using R
CA030302	- Exploratory Data Analytics for NLP
CA030303	- Computational Research Methodology
Elective	- Elective 1
CA030304	- Statistical Programming Lab using R
CA030305	- Mini Project II

Semester IV

CA030401	- Data Visualisation
Elective	- Elective 2
Elective	- Elective 3
CA030402	- Project
CA030403	- Comprehensive viva-voce

Elective Group I

CA850301	- Semantic Web and Web Scraping	- (Semester III)E1
CA850401	- Text Analytics	- (Semester IV)E2
CA850402	- Big Data Analytics and Artificial Intelligence	- (Semester IV)E3

Elective Group II

CA860301	- Social Media Mining	- (Semester III)E1
CA860401	- Business Intelligence	- (Semester IV)E2
CA860402	- Business Data Analytics	- (Semester IV)E3

Elective Group III

CA870301	- Sentiment Analytics	- (Semester III)E1
CA870401	- Internet of Things and Data Management	- (Semester IV)E2
CA870402	- Deep Learning	- (Semester IV)E3

*The Colleges shall select any one of the elective group and has to be intimated to the controller of examinations within two weeks of the commencement of third semester, the selection of courses from different elective groups is not permitted

7.   Scheme

Semester	Course Code	Course Name	Type of Course	Teaching Hrs/Week		Credit	Total Credit
				Theory	Practical		
I	CA030101	Statistics for Data Analytics	Core	4		4	20
	CA030102	Introduction to Data Analytics and Machine Learning	Core	4		4	
	CA030103	Advanced Operating Systems	Core	3		3	
	CA030104	Data Structure using C	Core	3		3	
	CA030105	Python Programming for Analytics	Core	3		3	
	CA030106	Python & Data Structure Lab	Core Lab I		8	3	
II	CA030201	Mathematics for Data Analytics	Core	4		4	20
	CA030202	Advanced Database Management system	Core	4		4	
	CA030203	Data Mining and Analytics	Core	3		3	
	CA030204	Programming with Java	Core	4		4	
	CA030205	Java & SQL Lab	Core Lab II		8	3	
	CA030206	Mini Project I	Core Mini		2	2	

		Project I					
III	CA030301	Statistical Modeling using R	Core	4		4	21
	CA030302	Exploratory Data Analytics for NLP	Core	4		4	
	CA030303	Computational Research Methodology	Core	4		4	
		Elective 1	Elective	3		3	
	CA030304	Statistical Programming Lab using R	Core Lab II		5	3	
	CA030305	Mini Project II	Core Mini Project II		5	3	
IV	CA030401	Data Visualisation	Core	5		4	19
		Elective 2	Elective	5		4	
		Elective 3	Elective	5		4	
	CA030402	Project	Core		10	5	
	CA030403	Comprehensive viva-voce	Core			2	

SEMESTER I

Semester	Course Code	Course Name	Type of Course	Teaching Hrs/Week		Credit	Total Credit
				Theory	Practical		
I	CA030101	Statistics for Data Analytics	Core	4		4	20
	CA030102	Introduction to Data Analytics and Machine Learning	Core	4		4	
	CA030103	Advanced Operating Systems	Core	3		3	
	CA030104	Data Structure using C	Core	3		3	
	CA030105	Python Programming for Analytics	Core	3		3	
	CA030106	Python & Data Structure Lab	Core Lab I		8	3	

CA030101 Statistics for Data Analytics

Instructional hours /week   : 4

Total instructional hours   : 72

Credits   : 4

Module 1: Theory of Probability

Basic terminology – Mathematical probability – Statistical probability – Axiomatic approach to probability – Some theorems on probability – Conditional probability – Multiplication theorem of probability – Bayes' theorem – Geometric probability

Module 2: **Descriptive Measures**

Frequency Distribution – Graphics representation of a frequency distribution – Averages – Arithmetic Mean, median, mode – Geometric Mean – Harmonic Mean – Dispersion – Measures of Dispersion – Coefficient of dispersion – Moments – Skewness – Kurtosis

Module 3: Probability distribution and hypothesis Testing

Distribution Function – Discrete random variables – Continuous random variable – Two dimensional random variables – Discrete uniform distribution – Bernoulli distribution – Binomial distribution – Geometric distribution – Normal distribution – Uniform distribution – Exponential distribution – Types of sampling – Parameter and statistic – Tests of significance – Procedure for testing of hypothesis

Module 4: Correlation and Regression analysis

Introduction - correlation and causation - types of correlation - Karl Pearson's coefficient of correlation-direct method of finding out correlation coefficient - calculation of correlation coefficient when change of scale and origin is made. Regression: introduction - regression equation of y on x -regression equation of x on y.

Module 5: Time series and data analysis life cycle

Some representative time series – Objectives of time series analysis – Approaches to time series analysis – Statistical techniques for analysing time series – Stationary time series – The time plot – Transformations – Analysing series that contain a trend – Analysing series that contain seasonal variation – Autocorrelation and the correlogram - Handling real data

Data analytics lifecycle overview:Key roles for a successful analytics project – Various phases of data analytics life cycle – Discovery – Data preparation – Model planning – Model building – Communicating the results - Operationalizing the results

Books of study

1.	Christopher Chatfield "The analysis of Time series An Introduction (Sixth Edition) "
2.	Data Science and Big Data Analytics by EMC Education Services, Wiley Publications
3.	 S. P. Gupta- "Statistical Methods", Sultan Chand & Sons.
4.	S C Gupta and V K Kapoor Fundamentals of Mathematical Statistics
5.	Trevor Hastie, Robert Tibshirani, Jerome Friedman " The elements of Statistical Learning"

CA030102 - Introduction to Data Analytics and Machine Learning

Instructional hours /week   : 4

Total instructional hours   : 72

Credits   : 4

Module 1: Introduction to Data Analytics

Introduction, Types of data, Quality of data, Data Preprocessing - Example applications - Data collection and management, Sources of data, Data collection, Exploring and fixing data, Data storage and management, Using multiple data sources, - Basic Statistical Descriptions of Data, Descriptive Statistics -- Exploratory Data analysis- Measuring Data Similarity and Dissimilarity- Graphical representation of data.

Module 2: Foundations of Learning

Components of learning – learning versus design-characteristics of machine learning – learning models – types of learning – training versus testing- Features – error measures -Supervised, Unsupervised and Reinforcement Learning.

Module 3: Supervised Learning

Regression: Linear Regression - Model representation for single variable, Single variable Cost Function, Gradient Decent for Linear Regression, Multivariable model representation, Multivariable cost function, Ridge Regression, Lasso Regression,

Classification: Logistic Regression - Problem of Overfitting, Regularization - Nearest neighbor models -- Decision Trees – Support Vector Machine, Kernels- Model Validation Approaches

Module 4: Unsupervised Learning:

Clustering: K means – clustering around medoids – hierarchical clustering – Ensemble learning - bagging and random forests – boosting -- Dimensionality reduction -- Principal Component Analysis, Linear Discriminant Analysis

Module 5: Artificial Neural Networks

Biological Neurons, Neural Networks Model representation, Intuition for Neural Networks, Multiclass classification, Cost Function, Back Propagation Algorithm, Weights initialization, Neural Network Training

Reference Books:

1. Han, Jiawei, Jian Pei, and Micheline Kamber, "Data mining: concepts and techniques", 3 rd Edition, Elsevier, 2011.
2. T. M. Mitchell, "Machine Learning", McGraw Hill, 2017.
3. K. P. Murphy, "Machine Learning: A probabilistic perspective", MIT Press, 2012.
4. Machine Learning, Tom M. Mitchell
5. Building Machine Learning Systems with Python, Richert & Coelho
6. Yoshua Bengio, "Learning Deep Architectures for AI", Now Publishers Inc (2009)

CA030103 - ADVANCED OPERATING SYSTEMS

Instructional hours /week : 3

Total instructional hours   : 54

Credits   : 3

Module 1

Computer system architecture – single processor systems, multiprocessor systems, clustered systems. Operating system operations - dual mode and multimode operation. Process management, Memory management, Storage management. Computing Environments- Traditional computing, Mobile computing, Distributed systems, Client Server computing, Peer-to-Peer computing, Virtualization, Cloud computing, Real-time embedded systems.

System structures - Operating system services, System calls, Types of system calls, Operating system structure-Simple structure, Layered approach, Microkernals , Modules, Hybrid systems.

Module 2

Process management - Process concept - Process state, PCB, Process Scheduling -Scheduling queues, Schedulers, Context switch, Operations on processes - creation, termination, Interprocess Communication- Shared memory systems , Message Passing systems.

Multithreaded Programming - Overview, Multithreading Models.

Process Scheduling – Basic Concepts, Scheduling criteria , Scheduling algorithms- FCFS, SJF, Priority scheduling, RR scheduling, Multilevel queue scheduling, Multilevel Feedback queue scheduling,

Module 3

Process Synchronization - The critical section problem- Peterson's Solution, Synchronization hardware, Mutex Locks, Semaphores, Monitors, Monitor usage

Deadlocks – System model, Deadlock characterisation, Methods for handling deadlocks, Deadlock prevention, Deadlock avoidance, Deadlock detection, Recovery from deadlock.

Module 4

Memory management- Memory management strategies - Basic hardware, Address binding, Logical Vs Physical address space, Dynamic loading, Dynamic linking and shared libraries , Swapping ,Contiguous memory allocation ,segmentation , Paging - Basic method , Hardware support, Protection, Shared pages.

Virtual memory management :- Demand paging - Basic concepts, Performance of demand paging, Page Replacement, Page Replacement algorithms - FIFO, Optimal page replacement, LRU page replacement.

Module 5

Case study - The Linux System - Features, Advantages,Linux history , Design Principles, Kernel Modules, Process Management, Scheduling - Process Scheduling, Real-time Scheduling , Virtual Memory , File Systems, Interprocess Communication, Security .

Various types of shells available in Linux - Comparison between various shells - Linux Commands for files and directories - cd, ls, cp ,rm, mkdir, rmdir, pwd, file, more, less . Creating and viewing files using cat.

Reference Text

1. Abraham Silberschatz, Galvin, Gange, Operating Ssystem Concepts, 9th Edition, Wiley Publishers.
2. Milan kovic, Operating Systems, Second Edition.
3. Official Red hat Linux Users Guide- Red hat, Wiley Dreamtech India.

4. Christopher Negus, Red Hat Linux Bible - 2005 Edition,Wiley Dreamtech India.
5. Yeswant Kanethkar, Unix Shell Programming,First Edition, BPB .

CA030104 - Data Structures Using C

Instructional hours /week : 3

Total instructional hours   : 54

Credits   : 3

Module 1

Introduction: Variables, Data types, Conditional and Loop Structures, Pointers. Static and dynamic memory allocation. Dynamic memory allocation and pointers, Memory allocation operators in C- malloc(), calloc(), free() and realloc(). User defined data types in C. Recursion, Recursive functions in C.

Concept of data structures, classification of data structures, Primitive and Non-primitive, Operations on data structures.

Introduction to algorithms, Performance analysis-Space complexity, Time complexity, Amortised complexity, asymptotic notations, Performance measurement.

Module 2

Arrays: Organization, Representation and implementation of arrays, examples. Implementation of Stacks and Queues, Circular Queues, Priority Queues, Double ended queues, Applications of stacks and queues.

Sorting and Searching techniques: Linear and Binary search, Selection sort, Merge sort, Simple insertion sort, Quick sort, Shell sort, Radix sort.

Module 3

Lists: Representation and implementation of singly linked list, Circular linked lists, doubly linked list, Linked list representation of stacks and queues, examples.

Dynamic storage management. Boundary tag system. Garbage collection and compaction.

Module 4

Trees: Representation and Implementation, Binary trees, insertion and deletion of nodes in binary tree, binary tree traversals, Binary search trees, Threaded Binary trees, Balanced trees (AVL trees), B- trees- Insertion and Deletion of nodes, Tree search

Module 5

Graphs: Directed Graphs, Shortest Path Problem, Undirected Graph, Spanning Trees, Techniques for graphs –Breadth First Search (BFS) and traversal, Depth First Search (DFS) and traversal

Hashing: Static hashing, hash tables, hash functions, overflow handling.

Reference Text

1. Ellis Horowitz, Sahni, Anderson-Freed, Fundamentals of Data Structures in C, Galgotia Publications
2. G S Baluja, Data structures Through C, Pearson
3. Aaron M. Tanenbaum, Data Structures Using C, Prentice Hall International
4. Ashok N. Kamthane, Introduction to data structures in C, Pearson

CA030105 - Python Programming For Analytics

Instructional hours /week : 3

Total instructional hours   : 54

Credits   : 3

Module 1

Structure of Python Program, Underlying mechanism of Module Execution-Branching and Looping-Problem Solving Using Branches and Loops-Functions – Lists and Mutability-Problem Solving Using Lists and Functions. Sequences, Mapping and Sets- Dictionaries- - Classes: Classes and Instances-Inheritance Exception Handling-Introduction to Regular Expressions using 're' module.

Module 2

The NumPy Library, Ndarray,Basic Operations ,Indexing, Slicing, and Iterating, Conditions and Boolean Arrays, Shape Manipulation, Array Manipulation,Structured Arrays, Reading and Writing Array Data on Files The pandas Library—An Introduction, Introduction to pandas Data Structures, Other Functionalities on Indexes, Operations between Data Structures, Function Application and Mapping, Sorting and Ranking,.

Module 3

 Introduction to Pandas Objects- Data indexing and Selection-Operating on Data in Pandas-Handling Missing Data-Hierarchical Indexing – Combining Data Sets. Aggregation and Grouping-Pivot TablesVectorized String Operations –Working with Time Series-High Performance Pandas- and query ()

Module 4

 Basic functions of matplotlib –Simple Line Plot, Scatter Plot-Density and Contour Plots-Histograms, Binnings and Density-Customizing Plot Legends, Colour Bars- Three-Dimensional Plotting in Matplotlib.

Module 5

Machine Learning with scikit-learn: The scikit-learn Library, Machine Learning :Supervised and Unsupervised Learning , Training Set and Testing Set, Supervised Learning with scikit-learn.

Reference Text :

1. Jake Vander Plas ,Python Data Science Handbook – Essential Tools for Working with Data, O'Reilly Media,Inc, 2016
2. Zhang.Y. , An Introduction to Python and Computer Programming, Springer Publications, 2016
3. Fabio Nelli , "Python Data Analytics Data Analysis and Science Using Pandas, matplotlib, and the Python Programming Language ", Apress, 2015
4. Wes McKinney, (2017) Python for Data Analysis: Data Wrangling with Pandas, NumPy, and Ipython, 2 nd Edition, O'Reilly Media.
5. Haslwanter, T.(2015) An Introduction to Statistics with Python, Springer

CA030106 – Python & Data Structure Lab

Instructional hours /week : 8

Total instructional hours   : 144

Credits   : 3

Data Structures Using C

1. Array implementation – Insertion of new element into a specified position, Deletion of an element from the specified position within the array
2. Stack implementation – PUSH, POP and Traverse
3. Queue implementation –Insertion, deletion and Traverse
4. Circular Queue implementation –Insertion, deletion and Traverse
5. Deque (Double ended queue) implementation –Insertion, deletion and Traverse
6. INFIX to POSTFIX Conversion
7. INFIX to PREFIX conversion
8. POSTFIX evaluation
9. Searching - Linear and Binary search using arrays
10. Sorting – Selection sort, Merge sort, Simple insertion sort, Quick sort, Shell sort, Radix sort
11. Lists implementation - Singly linked list, Circular linked list, Doubly linked list
12. Dynamic array implementation- Linked list representation and implementation of stack and queue operations
13. Creation of binary tree, counting no. of nodes and display the nodes in a tree
14. Searching a node in a binary tree
15. Insertion and deletion of nodes in a B-Tree
16. Graphs – Implementation of BFS and DFS

Python Programming Lab

1. Python syntax, functions, packages and libraries-
2. Types-Expressions
3. Variables-String Operations.
4. Python Data Structures: lists & Tuple –Sets -Dictionaries.

5. Programming Fundamentals: Conditions and Branching- Loops-Functions-Objects and Classes

6. Importing Datasets: Understanding the Dataset

7. Importing and Exporting Data in Python

8. Introduction to python libraries: Numpy- Scikit- Pandas-Matplotlib.-

9. Data cleansing and pre-processing: Identify and Handle Missing Values

10. Data Formatting

11. Data Normalization Sets

12. Regression Models: Linear Regression (SLR & MLR)

13. Logistic Regression

14. Decision Tree

15. K Nearest Neighbor- Random Forest

16. Gradient Boosting algorithms: XGboost

17. Support Vector Machine

18. Clustering Techniques: K means clustering

19. Apriori algorithm.

20. Model Evaluation: Over-fitting, Under-fitting

SEMESTER II

Semester	Course Code	Course Name	Type of Course	Teaching Hrs/Week		Credit	Total Credit
				Theory	Practical		
II	CA030201	Mathematics for Data Analytics	Core	4		4	20
	CA030202	Advanced Database Management System	Core	4		4	
	CA030203	Data Mining and Analytics	Core	3		3	
	CA030204	Programming with Java	Core	4		4	
	CA030205	Java & SQL Lab	Core Lab II		8	3	
	CA030206	Mini Project I	Core Mini Project I		2	2	

CA030201 – Mathematics for Data Analytics

Instructional hours /week : 4

Total instructional hours   : 72

Credits   : 4

Module 1

Mathematical Logic: Propositional Calculus: Statements and notations, Connectives: negation, conjunction, disjunction, statement formulas and truth tables, conditional and biconditional, Well-formed formulas, tautologies, equivalence of formulas, tautological implication. Normal forms: Disjunctive and conjunctive normal forms.

Predicate calculus: Predicates, statement functions, variables and quantifiers, predicate formulas, free & bound variables, universe of discourse.

Module 2

Set Theory- Sets, Set operations, Functions, Sequences and Summations

Module 3

Linear Algebra: Matrices and their properties (determinants, traces, rank, nullity, etc.); Eigenvalues and eigenvectors; Matrix factorizations; Inner products; Distance measures; Projections; Notion of hyperplanes; half-planes.

Module 4

Optimization: Unconstrained optimization; Necessary and sufficiency conditions for optima; Gradient descent methods; Constrained optimization, KKT conditions; Introduction to non-gradient techniques; Introduction to least squares optimization; Optimization view of machine learning.

Module 5

Fuzzy logic: Introduction, Crisp set an overview, Fuzzy sets basic types, Basic concepts, Characteristics and significance of paradigm shift.

Reference Text

1.   J.P. Tremblay & R Manohar- Discrete Mathematical Structures with Applications to Computer Science ,Mc Graw Hill.
2.   G. Strang (2016). Introduction to Linear Algebra, Wellesley-Cambridge Press, Fifth edition, USA.
3.   George J Klir & Bo Yuan- Fuzzy sets and Fuzzy logic Theory and applications, Prentice hall of India.
4.   David G. Luenberger (1969). Optimization by Vector Space Methods, John Wiley & Sons (NY)
5.   Kenneth H Rosen- Discrete Mathematics and its applications, Sixth Edition
6.   Edwin K P Chong and Stanislaw H Zak, An introduction to optimization , 4th Edition , Wiley

CA030202-- Advanced Database Management System

Instructional hours /week   : 4

Total instructional hours   : 72

Credits   : 4

Module 1

Database, need for DBMS, users, DBMS architecture, data models, views of data, data independence, database languages, Relational Model-Basic concepts, keys, integrity constraints, ER model-basic concepts, ER diagram, weak entity set, ER to Relational, relationships, generalization, aggregation, specialization

Module 2

Codd's rules, Relational model concepts , Relational algebra- Select, Project, Join, Relational calculus-tuple relational calculus and domain relational calculus, Specifying constraints management systems, Anomalies in a database, Functional dependencies, Normalization-First, Second, Third, Boyce Codd normal forms, multi-valued dependency and Fourth normal form, Join dependency and Fifth normal form.

Relational database query languages-Basics of SQL, Data definition in SQL- Data types, Creation, Insertion, Viewing, Updation, Deletion of tables, Modifying the structure of the tables, Renaming, Dropping of tables, Data constraints-I/O constraints, ALTER TABLE command.

Module 3

Database manipulation in SQL- Computations done on the table- Select command, Logical operators, Range searching, Pattern matching, Grouping data from tables in SQL, GROUP BY, HAVING clauses, Joins-Joining multiple tables, Joining tables to itself, DELETE, UPDATE, Views-Creation, Renaming the column of a view, Destroys view- Program with SQL, Security-locks, Types of locks, Levels of locks, Cursors - working with cursors, error handling, Developing stored procedures,-Creation, Statement blocks, Conditional execution, Repeated execution, Cursor-based repetition, Handling Error conditions, Implementing triggers, Creating triggers, Multiple trigger interaction.

Module 4

Concept of transaction, ACID properties, serializability, states of transaction, Concurrency control, Locking techniques, Time stamp based protocols, Granularity of data items, Deadlock, Failure classifications, storage structure, Recovery & atomicity, Log base recovery, Recovery with concurrent transactions, Database backup & recovery, Remote Backup System, Database security issues

Module 5

Object Oriented Database Management Systems (OODBMS) - concepts, need for OODBMS, composite objects, issues in OODBMSs, advantages and disadvantages of OODBMS. Distributed databases - motivation - distributed database concepts, types of distribution, architecture of distributed databases, the design of distributed databases, distributed transactions, commit protocols for distributed databases

Reference Text

1. Elmasri and Navathe, Fundamentals of Database Systems, 5th Edition, Pearson
2. Abraham Silbersehatz, Henry F. Korth and S.Sudarshan, Database System Concepts, 6 th Edition, Tata McGraw-Hill.

3. James R.Groff and Paul N. Weinberg The complte reference SQL Second edition,Tata McGraw Hill

CA030203-- Data Mining and Analytics

Instructional hours /week   : 3

Total instructional hours   : 54

Credits   : 3

Module 1

Introduction to Data mining, Data Mining Tasks, KDD process, Technologies for data mining, Application areas of data mining, Major issues in Data Mining, Data objects and Attribute types- Nominal, Binary, Ordinal and Numeric attributes, Measuring the central tendency-Mean, Median and Mode. Data Warehouse.

Module 2

Data Preprocessing: Needs of Pre-processing the Data, Data Cleaning- Missing Values, Noisy Data, Data Cleaning as a Process. Data Integration- Redundancy and correlation analysis, Data Reduction- Attribute Subset Selection, Dimensionality Reduction, Numerosity Reduction, PCA. Data Transformation strategies, Data transformation by Normalization, Discretization by Binning, Histogram Analysis

Module 3

Association Analysis- Frequent patterns, Basic terminology in association analysis- Binary representation, Itemset and support count, Association Rule, Support and Confidence, Frequent Item set generation- The Apriori Algorithm, Generating Association Rules from Frequent Itemsets, FP Growth algorithm, Pattern evaluation Methods. From Association Analysis to Correlation Analysis, Constraint-Based Frequent pattern Mining, Metarule-Guided Mining of Association Rules.

Module 4

Classification :- Basic concepts, General approach to classification, Decision Tree Induction, Basic Decision Tree algorithm, Attribute Selection Measures- Information Gain, Gain Ratio, Gini Index, Bayes Classification methods- Bayes' Theorem, Naïve Bayesian Classification, Rule-based Classification - Using IF-THEN Rules for Classification, Rule Extraction from a Decision Tree, Rule Induction Using a Sequential Covering Algorithm. Metrics for evaluating classifier performance, Cross validation. Classification by Back propagation- A Multilayer Feed-Forward Neural Network, Defining a Network Topology, Backpropagation.

Module 5

Cluster Analysis: Introduction, Basic Clustering methods- Partitioning methods- k-Means and k-Medoid. Hierarchical Methods - Agglomerative and Divisive Hierarchical Clustering. Density Based Methods - DBSCAN, OPTICS, DENCLUE. Grid Based- STING, CLIQUE, Outlier Analysis- what are outliers, Types of outliers, Outlier detection methods.

Reference Text

1. Jiawei Han & Micheline Kamber , Data Mining, Concepts and Techniques, , 3rd Edition.
2. Pang Ning Tan, Michael Steinbach and Vipin Kumar, Introduction to Data Mining, Pearson India Education Services
3. Arun K Pujari, Data Mining Techniques, , University Press

4. Sam Anahory & Dennis Murray, Data Warehousing in the Real World, Pearson Education, Asia.
5. Paulraj Ponnaiah, Data Warehousing Fundamentals, Wiley Student Edition

CA030204 - Programming with Java

Instructional hours /week   : 4

Total instructional hours   : 72

Credits   : 4

Module 1: **Object Oriented Programming Concepts and Basics of Java.**

Java Programming Environment – JDK, Java Virtual Machine, Bytecode, Features of Java Flow Control Statements – Conditional Statements, Iteration Statements, Jump Statements Arrays –One Dimensional Array, Multi-dimensional Array , Object Oriented Programming Concepts- (Objects and Classes, Encapsulation, Inheritance, Polymorphism) , Type of Inheritance , Method Overloading, Method Overriding, Dynamic Method Despatch

Module 2: **Input/Output Handling**

Constructors- Constructor Overloading , this, super, final, abstract and static Keywords, Interfaces- Defining an Interface, Implementing Interface, Extending Interfaces. String - String Handling Fundamentals, Comparison of String and StringBuffer Class, Special String Operations- Character Extraction, String Comparison, Searching String, Modifying a String, String Copy ,Input and Output Streams – Byte Stream , Character Stream

Module 3: **Packages; Exception Handling and Thread**

Packages – Defining Packages, Built in Packages (java.lang, java.util, java.io, java.net, javax.swing), Importing Packages, Implementation of User Defined Packages, Access Protection in Java, Exception Handling - try, catch, throw, throws and finally Statements, Java's Built-in Exceptions, Creating User Defined Exceptions. Threads- Thread Lifecycle, Thread Priorities, The Thread Class, Runnable Interface, Creating a Thread – Implementing Runnable, Extending Thread, Inter Thread Communication, Suspending Resuming and Stopping Threads.

Module 4: **GUI Programming**

 Basic Event Handling – Delegation Event Model, Important Event Classes And Listener Interfaces, Handling Mouse and Keyboard Events, Adapter Classes, Swing -Window Fundamentals – Class Hierarchy, Frame, Creating a Simple Window Based Application, ImageIcon, JLabel, JTextField, JTextArea, JButton, JCheckBox, JRadioButton, JList, JComboBox, JTable, JTabbedPane, JScrollPane, Layout Management – The FlowLayout, BorderLayout, GridLayout, CardLayout

Module 5: **File, Database and RMI**

File Management - Reading and Writing Files (FileInputStream and FileOutputStream Classes), JDBC – Components of JDBC, JDBC architecture, various kinds of JDBC drivers, The Structured Query Language, The Connection Interface, The Statement Interface, The

PreparedStatement Interface, Scrollable and Updatable ResultSets, RowSets, Transactions. Remote Method Invocation (RMI) – Client Server Application using RMI.

Reference Text

1. Herbert Schildt Java 2 The Complete Reference, Tata McGraw Hill (5th Edn.)
2. DT Editorial Services, Java 8 Programming Black Book, Dreamtech Press.
3. James. P. Cohoon, Programming java 5.0, Jack. W. Davison (Tata McGraw Hill)
4. C Thomas Wu, An introduction to Object Oriented Programming with Java, , Tata McGraw Hill, (2006)
5. Wigglesworth and McMillan, Java Programming: Advanced Topics, , Cengage Learning India, 3rd Edn.
6. Bernard Van Haecke, JDBC: Java Database Connectivity, , IDG Books India (2000)

CA030205 –Java and SQL Lab

Instructional hours /week : 8

Total instructional hours : 144

Credits : 3

Advanced Java Programming

1. Basic Concepts and File Handling
1.1. Inheritance, Polymorphism
1.2. Constructors
1.3. Interface
1.4. Package
1.5. One Dimensional and Two Dimensional Array Manipulation
1.6. String Handling (Character Extraction, String Comparison, Searching String, Modifying a String, String Copy)
1.7. Exception (Built-in and User Defined)
1.8. Thread (Using Runnable Interface and Thread Class)
1.9. File management (File reading, Writing, Appending and Content Replacing)
2. GUI, Database and RMI
2.1. Event Handling (Keyboard and Mouse Events)
2.2. Working with Swing (ImageIcon, JTextField, JTextArea, JButton, JCheckBox, JRadioButton, JComboBox, JList, JTable)
2.3. Layout Management (The FlowLayout, BorderLayout, GridLayout, CardLayout)
2.4. Simple Programs of Database Connectivity
2.5. Demo Client Server Application using RMI

SQL

1. Creating database tables and using data types (create table, modify table, drop table).
2. Data Manipulation (adding data with INSERT, modify data with UPDATE, deleting

3. records with DELETE).
4. Implementing the Constraints (NULL and NOT NULL, primary key and foreign key Constraint, unique, check and default constraint).
5. Retrieving Data Using SELECT (simple SELECT, WHERE, IN, BETWEEN, ORDERED BY, DISTINCT and GROUP BY).
6. Aggregate Functions (AVG, COUNT, MAX, MIN, SUM).
7. String functions.
8. Date and Time Functions.
9. Use of union, intersection, set difference.
10. Implement Nested Queries & JOIN operation.
11. Performing different operations on a view.
12. Stored Procedure Programming – Simple Procedures – decision making – Loops – Error handlers – Cursors – Functions - Triggers – Calling Stored Procedure from Triggers.

CA030206 Mini Project I

Instructional hours /week : 2

Total instructional hours   : 36

Credits   : 2

Mini Project aims at giving students hands-on experience in applying the programming knowledge in python to solve a real-world situation/problem using techniques in Data mining and Machine learning. Students must take up individual project. Evaluation of the project is internal.

SEMESTER III

Semester	Course Code	Course Name	Type of Course	Teaching Hrs/Week		Credit
				Theory	Practical	
III	CA030301	Statistical Modeling using R	Core	4		4
	CA030302	Exploratory Data Analytics for NLP	Core	4		4
	CA030303	Computational Research Methodology	Core	4		4
		Elective 1	Elective	3		3
	CA030304	Statistical Programming Lab using R	Core Lab II		5	3
	CA030305	Mini project II	Core Mini Project II		5	3

CA030301 - Statistical Modeling using R

Instructional hours /week   : 4

Total instructional hours   : 72

Credits   : 4

Module 1: **Introduction to Statistical Modelling**

Introduction: Statistics in Real World – Models in data analysis - Populations and samples- Variables and factors - Observational and experimental Data - Collecting Good Quality Data- Statistical models.

Module 2: **R : An Introduction**

Introduction to R - Work spaces - R Syntax and programming - Variables - Operators, Vectors - List - Matrices - Arrays, Factors, Data Frames - Values Reading data using R - Basic read write operations.

Module 3: **R Programming**

Conditions and Loops - Functions: Call & Write - Exception Handling - Progress & Timings - Visibility

Module 4: **Elementary Statistics**

Describing Raw Data - Sampling Distributions - Confidence Intervals - Hypothesis Testing - Components - Testing Means - Testing Proportions & Categorical Variables - General Concepts of Simple Linear Regression.

Module 5: **Statistical Testing And Modeling**

Sampling Distributions and Confidence - Hypothesis Testing - Analysis of Variance - Simple Linear Regression - Multiple Linear Regression - Linear Model Selection and Diagnostics.

Reference Text

1. Tilman M. Davies " The Book of R" 2016, No Starch Press
2. W. J. Krzanowski " An Introduction to Statistical Modelling" Wiley
3. Andy Field , Jeremy Miles, Zoe Field "Discovering Statistics Using R" , April 5th 2012 by Sage Publications Ltd
4. Joseph Schmuller , " Statistical Analysis with R For Dummies" May 2017, Wiley

CA030302—Exploratory Data Analytics for NLP

Instructional hours /week : 4

Total instructional hours   : 72

Credits   : 4

Module I
Introduction-Natural language processing and its neighbors-Three themes in natural language-processing-Learning and knowledge-Search and learning-Relational, compositional, and distributional perspectives.

Module II
Language Processing and Python -Computing with Language: Texts and Words-Texts as Lists of Words-Computing with Language: Simple Statistics-Making Decisions and Taking Control-Automatic Natural Language Understanding.

Module III
Fundamentals of Exploratory Data Analysis-Significance of EDA-Making sense of Data-Software tools-Getting started with EDA-Numpy,Pandas,SciPy,Matplotlib.
Visual Aids for EDA-Line Chart,Bar Chart,Scatter Plot,Pie Chart, Table Chart, Polar Chart,Histograms,Lollipop Chart-Choosing the best chart.

Module IV
Data Transformation-Background-Merging database-style dataframes-Transformation Techniques-Data duplication-Replacing values-Handling missing data-Renaming axes indexes-Discretization and binning-Outlier detection and filtering-Benefits of data transformation.
Grouping Datasets-Groupby mechanics-Data aggregation-Pivot tables.

Module V

Hypothesis testing and regression- Hypothesis testing-p-hacking-Types of regression-Constructing a linear regression model-Implementing a multiple linear regression model

Model Development and Evaluation-Supervised and unsupervised learning-Reinforcement learning-Machine Learning Workflow.

Reference Text

1. Natural Language Processing by Jacob Eisenstein
2. Natural Language Processing with Python by Steven Bird, Ewan Klein, Edward Lopper
3. Hands-On Exploratory Data Analysis with Python by Suresh Kumar Mukhiya, Usman Ahmed

CA030303— Computational Research Methodology

Instructional hours /week : 4

Total instructional hours : 72

Credits   : 4

Module 1: **Research Methodology**

Research Methodology: Meaning of Research, Objectives of Research, Motivation in Research, Types of Research, Research Approaches, Significance of Research, Research Methods versus Methodology, Research and Scientific Method, Research Process, Reading and Reviewing-Research literature, Finding Research Papers, Critical Reading, Developing a literature Review, Guidelines for Research Skills and Awareness, Validity of Research,Criteria of Good Research.

Module 2: **Data in Research**

Introduction, Need for Data Collection, Methods of Data Collection, Principles for Accessing Research Data, Data Processing, Data Analysis, Presentation of Data, Error Analysis, Scientific Models. Scientific Methodology - Introduction Rules and Principles of Scientific Method, Hypothesis,Testing of Hypothesis, Basic concepts, Procedure,Important parametric tests: z-test ,t-test, $\chi 2$ -square test, F test

Module 3: **Research Methods for Computer Science**

History of ideas in computing, Evolution of Computing Research, Overview of Research Methods: Measurements based research methods - Deductive Methods - Inductive Methods - The significance of Interdisciplinary research

Module 4: **Ethics in Research**

Ethics in Research –Research Ethics, Importance of Ethics in Research, Ethics values and Principles, Some Ethical issues ,Plagiarism, Misuse of Privileged Information, Misuse of Data, Authorship and other publication issues, Copy Right, Intellectual Property Rights.

Module 5: **Reporting and Thesis Writing**

Presentation of algorithms, Environment of Algorithms, Asymptotic Cost. Graphs. Technical Reports- Structuring General format, Report-Bibliography referencing and footnotes.Research

in Practice- Literature Review, Journals, Conference Proceedings, journal Impact Factor, citation Index, h Index .Application of Computer in Research --MS office and its application in Research, Use of Internet in Research – Websites, search Engines, E-journal and E-Library.

Reference Text

1. Kothari, C.R., 1990. Research Methodology: Methods and Techniques. New Age International. Publishers(Second revised edition)
2. Krishnan Nallaperumal, "Engineering Research Methodology : A Computer Science and Engineering and Information and Communication Technologies Perspective. " (First Edition)
3. Justin Zobel,Writing For Computer Science, Springer (Third Edition)
4. K Prathapan,Research Methodology for Scientific Writing ,I.K International Publishing House Pvt.Ltd
5. S.P Satarkar, S.V., 2000. Intellectual Property Rights and Copy right. Ess Publications

ELECTIVES

CA850301— Semantic Web and Web Scraping

Instructional hours /week : 3

Total instructional hours   : 72

Credits   : 3

Module 1: **Introduction to Semantic Web**

Web 2.0 and 3.0 – Meaning of Semantic Data – Distributed web of data – Metadata - Features of semantic web – Data across the web – The basics of semantic web - The Limits of the web – The vision of the semantic web – Semantic web standards – RDF – RDF Scheme (RDFS) – OWL Web Ontology Language – SPARQL Protocol – RDF Query Language (SPARQL) - Need of RDFS – Machine Readability – core elements of RDFS – XML Schema – RDF schema

Module 2: **Knowledge Representation Methods**

The concept of Ontology - SKOS – Representation of thesauri - Glossaries – Scheme classification – Taxonomies – Controlled Vocabularies - Hierarchical Structure – Formal Representations - Standard Namespaces – JSON based serialization for Linked Data - RDF Triple stores – Turtle – RDFa – Internal Identifiers - URI – RDFS – Classes – Resources – Inferred Property Characterization – Literals – Linked Open Data – DBpedia – Querying RDF Graphs – Vocabularies – Graph based validation - Shape constraint Language (SHACL)

Module 3: **Data Integration**

Data Integration issues- Data Interoperability – Data Migration – Data Representation Formats – Data Silos – Linked Data Management – Knowledge Mining Stack – NLP – Named Entity Recognition – Machine Learning – Knowledge Graphs

Module 4: **Use cases & Contemporary Issues**

Use cases Specifications and Discussion: - Transportation: Data Sources – Representation – Linked Data Mapping - Knowledge Modeling – Telecommunication – Knowledge Modeling – Customer Care Support Documents – Internal Reports – Named Entity Recognition – Linked Data Mapping.

Contemporary Issues : Customer provider mismatch – Interlinking domain specific information – Combining different services from different providers – contrast with contemporary web applications Markup languages – Object Access Protocols – Service description – Discovery – Integration

Module 5: **Web Scraping**

Introduction - Dealing with HTTP & Java Script - Web Scrawling - Managerial and Legal Concerns.

Reference Text

1. Dean Allemang, James Hendler: "Semantic Web for the Working Ontologist Effective Modeling in RDFs and OWL", 2nd Edition, 2008.
2. Seppe vanden Broucke, Bart Baesens "Practical Web Scraping for Data Science: Best Practices and Examples with Python", Apress
3. Liyang Yu, "Introduction to the Semantic Web and Semantic web services" Chapman & Hall/CRC, Taylor & Francis group, 2007.
4. Toby Segaran, Colin Evans, Jamie Taylor, "Programming the Semantic Web", 1st Edition, July 2009.
5. Pollock, J.T.: Semantic web for dummies. Wiley Publishing, Inc., Indianapolis, 2009.

CA860301— Social Media Mining

Instructional hours /week : 4

Total instructional hours   : 72

Credits   : 3

Module 1:

Introduction-New Challenges for Mining, Graph basics- Graph Representation , Types of Graphs, Connectivity in Graphs, Special Graphs, graph algorithms, Network measures centrality, transitivity and reciprocity, balance and status, similarity, Network Models - Properties of Real-World Networks, Random Graphs, Small-World Model , Preferential Attachment Model

Module 2:

Data Mining Essentials- Data, Data Preprocessing, Data Mining Algorithms, Supervised Learning , Unsupervised Learning.

Module 3:

Communities and Interactions- Community Analysis, Community Evolution, Community Evaluation Information Diffusion in Social Media- Herd Behavior, Information Cascades , Diffusion of Epidemics.

Module 4:

Influence and Homophily- MeasuringAssortativity , Influence, Homophily , Distinguishing Influence and Homophily

Recommendation in Social Media- Challenges , Classical Recommendation Algorithms, Recommendation Using Social , Evaluating Recommendations

Module 5:

Behavior Analytics- Individual Behavior, Individual Behavior Analysis, Individual Behavior Modelling, Individual Behavior Prediction, Collective Behavior

Reference Text

1. Reza Zafarani, Mohammad Ali Abbasi. Huan, " Social Media Mining- An Introduction", Cambridge University Press, 2014.
2. Jure Leskovec,AnandRajaraman, Jeffrey D. Ullman, "SMining of Massive Datasets".

CA870301— Sentiment Analytics

Instructional hours /week : 3

Total instructional hours : 72

Credits : 3

Module 1: **Introduction to Sentiment Analysis**

Introduction: Sentiment Analysis Applications - Sentiment Analysis Research - Sentiment Analysis as Mini NLP. The Problem of Sentiment Analysis: Definition of Opinion - Definition of Opinion Summary - Affect, Emotion, and Mood - Different Types of Opinions - Author and Reader Standpoint. Document Sentiment Classification: Supervised Sentiment Classification - Unsupervised Sentiment Classification - Sentiment Rating Prediction - Cross-Domain Sentiment Classification - Cross-Language Sentiment Classification - Emotion Classification of Documents.

Module 2: **Subjectivity Classification and Challenges**

Sentence Subjectivity and Sentiment Classification: Subjectivity - Sentence 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 - Emotion Classification of Sentences.

Module 3: **Aspect Oriented Classification**

Aspect Sentiment Classification: - Rules of Sentiment Composition - Negation and Sentiment - Modality and Sentiment - Coordinating Conjunction But - Sentiment Words in Non-opinion Contexts - Rule Representation - Word Sense Disambiguation and Co reference Resolution. Aspect and Entity Extraction: Frequency-Based Aspect Extraction - Exploiting Syntactic

Relations - Using Supervised Learning - Mapping Implicit Aspects - Grouping Aspects into Categories - Exploiting Topic Models - Entity Extraction and Resolution - Opinion Holder and Time Extraction.

Module 4: **Sentiment Lexicon generation and Summarization**

Sentiment Lexicon Generation: Dictionary-Based Approach - Corpus-Based Approach - Desirable and Undesirable Facts. Analysis of Comparative Opinions: Problem Definition - Identify Comparative Sentences - Identifying the Preferred Entity Set - Special Types of Comparison - Entity and Aspect Extraction. Opinion Summarization and Search: Aspect-Based Opinion Summarisation - Enhancements to Aspect-Based Summary - Contrastive View Summarization - Traditional Summarization - Summarization of Comparative Opinions - Opinion Search - Existing Opinion Retrieval Techniques. Mining Intentions: Problem of Intention Mining - Intention Classification - Fine-Grained Mining of Intentions.

Module 5: **Identifying intention, fake and quality of opinion**

Detecting Fake or Deceptive Opinions: Different Types of Spam - Supervised Fake Review Detection - Supervised Yelp Data Experiment - Automated Discovery of Abnormal Patterns - Model Based Behavioral Analysis - Group Spam Detection - Identifying Reviewers with Multiple User ids - Exploiting Business in Reviews - Some Future Research Directions. Quality of Reviews: Quality Prediction as a Regression Problem - Other Methods - Some New Frontiers.

Reference Text

1. Bing Liu "Sentiment Analysis: Mining Opinions, Sentiments and Emotions, Cambridge University Press, 2015.
2. Bing Liu "Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers, 2012.
3. Erik Cambria, Dipankar Das "A Practical Guide to Sentiment Analysis" Springer, 2017.

CA030304 –Statistical Programming Lab using R

Instructional hours /week : 5

Total instructional hours   : 90

Credits   : 3

Statistical Programming Lab using R

Cycle 0 / Prerequisite : Introduction to R

- Work spaces
- Variables
- Operators
- Vectors

- List

- Matrices

- Arrays

- Factors

Cycle1: Programming with R

- Set Conditions and Loops

- Create Functions and use them

- Try Exception Handling

Cycle 2: Getting Used to R: Describing Data

- Viewing and Manipulating Data
- Plotting Data
- Reading in Your Own Data

Cycle 3: Visualizing Data

- Tables, charts and plots.
- How to find the mean median standard deviation and quantiles of a set of observations.
- Students may use real as well as artificial data sets

Cycle 4: Probability Distributions.

- Generate and Visualize Discrete and continuous distributions using the statistical environment. Demonstration of CDF and PDF uniform and normal, binomial Poisson distributions.
- Students are expected to generate artificial data using and explore various distribution and its properties. Various parameter changes may be studied.

Cycle 5: Densities of Random Variables

- Off the Shelf Distributions in R
- Matching a Density to Data
- More About Making Histograms

Cycle 6: Binomial Distribution

- Study of binomial distribution. Plots of density and distribution functions. Normal
- approximation to the Binomial distribution.

Cycle 7: Building Confidence in Confidence Intervals

- Populations Versus Samples
- Large Sample Confidence Intervals
- Simulating Data Sets
- Evaluating the Coverage of Confidence Intervals

Cycle 8: Perform Tests of Hypotheses

- How to perform tests of hypotheses about the mean when the variance is known.
- How to compute the p-value.

- Explore the connection between the critical region, the test statistic, and the p-value

Cycle 9: Correlation

- How to calculate the correlation between two variables.
- How to make scatter plots.
- Use the scatter plot to investigate the relationship between two variables

Cycle 10: Estimating a Linear Relationship

- A Statistical Model for a Linear Relationship
- Least Squares Estimates
- The R Function lm
- Scrutinizing the Residuals

CA030305 Mini Project II

Instructional hours /week : 5

Total instructional hours   : 90

Credits   : 3

Mini Project aims at giving students hands-on experience in applying the programming knowledge in python to develop a real application for data analytics. Students must take up individual projects. Evaluation of the project is external.

SEMESTER IV

Semester	Course Code	Course Name	Type of Course	Teaching Hrs/Week		Credit	Total Credit
				Theory	Practical		
IV	CA030401	Data Visualisation	Core	5		4	19
		Elective 2	Elective	5		4	
		Elective 3	Elective	5		4	
	CA030402	Project	Core		10	5	
	CA030403	Comprehensive viva-voce	Core			2	

CA030401— Data Visualisation

Instructional hours /week : 5

Total instructional hours   : 90

Credits   : 4

Module 1: **Introduction to Data Visualisation**

Definition – Methodology – Seven Stages of Data Visualisation - Data Visualisation Tools. Visualising Data: Mapping Data onto Aesthetics – Visualising Amounts - Visualising Distributions: Histograms and Density Plots – Visualising Propositions: – Visualising Associations: Among Two or More Quantitative Variables – Visualising Time Series and Other Functions of an Independent Variable – Trends – Visualising Geospatial Data.

Module 2: **R : Interactive Data Visualisation**

Introduction to D3 - Fundamental Technology: The Web – HTML – DOM – CSS – JavaScript – SVG. D3 Setup – Generating Page Elements – Binding Data - Drawing with data – Scales: Domains and Ranges – Normalization – Creating a Scale – Scaling the Scatter Plot – Other Methods and Other Scales. Axes – Modernizing the Chart – Update the Data – Transition – Updates – Interactivity.

Module 3: **D3 Based Reusable Chart Library**

Setup and Deployment – Generate Chart – Customize Chart: Additional Axis – Show Axis Label – Change Chart Type – Format Values – Size – Color – Padding –Tooltip. Use APIs: Load and Unload – Show and Hide – Focus – Transform – Groups – Grid – Regions – Flow

– Revert – Toggle –Legend – Sub chart – Zoom – Resize. Customize Style. Building Real time and Live Updating animated graphs with C3.

Module 4: **Tableau Introduction**

Environment Setup – Navigation – File & Data Types. TA SOURCE: Custom Data View – Extracting Data – Fields Operations – Editing Meta Data – Data Joining – Data Blending. Worksheets.

Module 5: **Basic and Advanced Charts in Tableau**

Bar Chart – Line Chart – Pie Chart – Scatter Plot – Bubble Chart –Gantt Chart – Histograms - Waterfall Charts. Dashboard – Formatting – Forecasting – Trend Lines.

Reference Text

1. Ben Fry, "Visualizing Data: Exploring and Explaining Data with the Processing Environment", O'Reilly, 1st Edition, 2008.
2. Scott Murray, "Interactive data visualization for the web: An Introduction to Designing with D3", O'Reilly, 2nd Edition, 2017.
3. Joshua N. Milligan, "Learning Tableau 2019: Tools for Business Intelligence, data prep, and visual analytics", Packt Publishing Limited, 2019.
4. Claus O. Wilke, "Fundamentals of Data Visualization: A Primer on Making Informative and Compelling Figures", O.Reilly, 2019.

ELECTIVES

CA850401— Text Analytics

Instructional hours /week : 5

Total instructional hours : 90

Credits : 4

Module 1: **Text Mining**

Text Mining - Definition - General Architecture – Core Text mining Operations. Nature of unstructured and semi-structured text, Collecting documents NLP : Text pre-processing Sentence Segmentation tokenization - lemmatization - stemming - Parsing text - keywords,-POS, Bag of Words Model, n-grams, chunking and Named Entity Recognition (NER) Corpus - sentence boundary determination - Textual information to numerical vectors -vector generation for prediction- document standardization and Representation – Inverted Index-term document matrix (TDM)-TDM Frequency

Module 2: **Information retrieval and Extraction**

Information retrieval- keyword search - Vector space scoring, Models - web- based document search-matching-inverted lists. Information extraction-Architecture - Co-reference - Named Entity and Relation Extraction-Template filling and database construction –Applications.

Inductive - Unsupervised Algorithms for Information Extraction. Text Summarization Techniques -Topic Representation -Influence of Context -Indicator representations.

Module 3: **Text Categorisation**

Text Categorisation – Definition – knowledge engineeringTextClassificationFeature Selection for Text Classification, Gini Index, Information Gain .Evaluating model : confusion matrix, class   specific measure Classification models : Decision Tree Classifiers -Rule- based Classifiers - Naive Bayes Classifiers - Linear Classifiers-Classification of Linked and Web Data –Text Clustering – Definition- Feature Selection and Transformation Methods for Text Clustering –Distance and   similarities-Hierarchical cluster –K-means -Semi- Supervised Clustering -Transfer LearningPattern Extraction - Apriori Algorithm – FP Tree algorithm - Results summaries.

Module 4: **Probabilistic Models for Text Mining**

Probabilistic Models: Introduction, Mixture Models, Stochastic Processes in Bayesian Nonparametric Models, Graphical Models,Probabilistic Models with Constraints, Parallel Learning Algorithms.Probabilistic Models for Information Extraction -Hidden Markov Models - Stochastic Context-Free Grammars - Maximal Entropy Modeling -Maximal Entropy Markov Models - Conditional Random Fields.

Module 5: **Text Analytics Use Cases**

Text Analytics in Social Media, Modeling text sentiments, Spam Detection, Mining Text Streams, Opinion Mining and Sentiment Analysis, Text Visualization Approaches - Architectural Considerations –Common Visualization Approaches for text mining.

Reference Text

1. Murugan Anandarajan "Practical Text Analytics: Maximizing the Value of Text Data", Springer; 2018
2. Charu C. Aggarwal, "Machine Learning for Text" Springer, 2018.
3. Steven Bird, Ewan Klein and Edward Loper"Natural Language Processing with Python"

CA860401— Business Intelligence

Instructional hours /week : 5

Total instructional hours   : 90

Credits   : 4

Module 1:

Business intelligence: Introduction, Concepts and Applications. Business Intelligence for better decisions, Business Intelligence types, tools, skills, applications. Decision making and

Analytics - business intelligence, analytics and decision support. Foundation and technologies for decision making.

Module 2:
Business applications of Decision Trees, Regression, Artificial Neural Networks, Cluster Analysis, Association Rule Mining - Techniques, Algorithm, Exercise, Advantages and Disadvantages.

Module 3:
Big data and future directions for Business Analytics- Big Data Analytics, Business Analytics ,Emerging Trends and Future Impacts. Business applications of Big Data, Technologies and Management Big data.

Module 4:
Predictive Analytics: Data mining in Business Intelligence- Text Mining,Web Mining - Business applications, practices and algorithms. Descriptive Analytics - Data warehousing, Business Reporting, Visual Analytics and Business Performance Management.

Module 5:

Understanding BI and Mobility , BI and Cloud Computing , Business Intelligence for ERP Systems, Social CRM and BI

Reference Text

1. Business Intelligence and Data Mining ,Anil K. Maheshwari, PhD,Business Expert Press, LLC,2015
2. BUSINESS INTELLIGENCE AND ANALYTICS: SYSTEM FOR DECISION SUPPORT,Ramesh Sharda(Oklahoma State University), Dursun Delen(Oklahoma State University),Efraim Turban(University of Hawaii), Pearson Education,Inc., 2015.10th Edition
3. Fundamentals of Business Analytics, 2ed,R N Prasad, Seema Acharya

CA870401— Internet of Things and Data Management

Instructional hours /week : 5

Total instructional hours   : 90

Credits   : 4

Module 1: **Introduction to IoT**

Introduction to IoT - Enabling technologies of IoT - IoT Architectures - IoT Data Management and Analytics - IoT levels and deployment templates – Application domains of IoT.

Module 2: **Data and Knowledge Management in IoT**

Stream Processing for IoT : Introduction - Foundations - Continuous logic processing systems - challenges and future.

Distributed Data Analytics for IoT : Introduction - Anomaly Detection Types - Incremental Logical Modeling

Module 3: **Cloud Computing & Fog Computing : IoT perspective**

Cloud Computing : Introduction, Architecture, Managing IoT resources, Framework for intelligent applications

Fog Computing : Introduction, Characteristics, Architecture, Application, Research

Module 4: **IoT Enablers and Solutions**

Introduction, Embedded Device Programming Languages, Message Passing in Devices, Coordination Languages, Polyglot Programming, IoT programming Approaches, Existing IoT frameworks, Virtualisation and Virtual Machines.

Module 5: **IoT Challenges & Applications**

Governing Internet of Things: Issues, Approaches, and New Paradigms, Obfuscation and Diversification for Securing the Internet of Things (IoT), Applications of IoT

Reference Text

1. Rajkumar Buyya, Amir Vahid Dastjerdi, "Internet of Things Principles and Paradigms", First Edition, Morgan Kaufman, 2016.
2. ArshdeepBahga, Vijay Madisetti, „Internet of Things: A Hands-On Approach", Universities Press, 2015.
3. Michael Margolis, "Arduino Cookbook" 2nd Edition, O'Reilly Media, 2012.
4. Charles Platt, "Make Electronics – Learning by discovery", O'Reilly Media, 2015.

CA850402— Big Data Analytics and Artificial Intelligence

Instructional hours /week : 5

Total instructional hours : 90

Credits   : 4

Module 1: **Introduction to Big Data**

Introduction to Big Data: - Evolution – Importance – Sources – Myths - Characteristics– Data Science Components - Usecases – Data Processing Models – Limitations of Conventional Data Processing Approaches - Data Discovery Traditional Approach, Big Data Exploration - Data Augmentation – Operational Analysis – Security and Intelligence – Data Analytics Classifications - Descriptive – Diagnostic - Predictive – Prescriptive – Augmented – Pervasive Analytics.

Module 2: **Big Data Technologies**

Big Data Technologies - Hadoop: Architecture - HDFS - MAPREDUCE : Frame work, PIG: Features - Modes - commands- functions, HIVE: Features - Architecture - Data Units, SPARK:

Advantages - Features - SPARK Ecosystem, SCALA Framework, NoSQL : Ideology & Databases - HBASE: Architecture- Managing large data sets with HBase,

Module 3: **Introduction to Artificial Intelligence**

Introduction to Artificial Intelligence: What is AI?, Importance of AI, – Knowledge: Definition and importance, Knowledge-based concepts, Knowledge Representation schemes: Formalised Symbolic Logics - Dealing with Inconsistencies and Uncertainties– Probabilistic Reasoning-

Module 4: **Knowledge Organization and Manipulation**

Search: Uninformed and informed - Graph Search, Matching Techniques - Structures - Measures - Patterns - Matching Algorithms, Knowledge Organization and Management - Indexing and Retrieval Technique - Knowledge Integration - Memory Organization System.

Module 5: **Expert Systems**

Experts Systems: Overview of an Expert System, Structure of an Expert Systems, Different Types of Expert Systems-Rule Based, Model Based, Case Based and Hybrid Expert Systems, Knowledge Ac - quisition and Validation Techniques, Black Board Architecture, Knowledge Building System Tools, Expert System Shells

Reference Text

1. V. Bhuvaneswari, T. Devi, "Big Data Analytics: A Practitioner's Approach", Sci-Tech Publications, 2016
2. D. W. Patterson, Introduction to Artificial Intelligence and Expert system, Prentice Hall of India
3. Field Cady, "The Data Science Handbook", John Wiley & Sons, Publishers, ISBN: 9781119092940, 2017
4. Elainc Rich, "Artificial Intelligence", Mc Graw Hill.

CA860402— Business Data Analytics

Instructional hours /week   : 5

Total instructional hours   : 90

Credits   : 4

Module 1: Business Analytics Overview
Introduction – Drivers for Business Analytics – Applications of Business Analytics: Marketing and Sales, Human Resource, Healthcare, Product Design, Service Design, Customer Service and Support – Skills Required for a Business Analyst – Framework for Business Analytics Life Cycle for Business Analytics Process.

Module 2: Business Analytics Essentials

Descriptive Statistics – Using Data – Types of Data – Data Distribution Metrics: Frequency, Mean, Median, Mode, Range, Variance, Standard Deviation, Percentile, Quartile, z-Score, Covariance, Correlation – Data Visualization: Tables, Charts, Line Charts, Bar and Column Chart, Bubble Chart, Heat Map – Data Dashboards.

Module 3: Modeling Uncertainty And Statistical Inference
Modeling Uncertainty: Events and Probabilities – Conditional Probability – Random Variables – Discrete Probability Distributions – Continuous Probability Distribution – Statistical Inference: Data Sampling – Selecting a Sample – Point Estimation – Sampling Distributions – Interval Estimation – Hypothesis Testing.

Module 4:Analytics Using Hadoop And Mapreduce Framework
Introducing Hadoop – RDBMS versus Hadoop – Hadoop Overview – HDFS (Hadoop Distributed File System) – Processing Data with Hadoop – Introduction to MapReduce – Features of MapReduce – Algorithms Using Map-Reduce: Matrix-Vector Multiplication, Relational Algebra Operations, Grouping and Aggregation – Extensions to MapReduce.

Module 5:Other Data Analytical Frameworks

Overview of Application development Languages for Hadoop – PigLatin – Hive – Hive Query Language (HQL) – Introduction to Pentaho, JAQL – Introduction to Apache: Sqoop, Drill and Spark, Cloudera Impala – Introduction to NoSQL Databases – Hbase and MongoDB.

Reference Text

1. VigneshPrajapati, ―Big Data Analytics with R and Hadoop‖, Packt Publishing, 2013.
2. Umesh R Hodeghatta, UmeshaNayak, ―Business Analytics Using R – A Practical Approach‖, Apress, 2017.
3. AnandRajaraman, Jeffrey David Ullman, ―Mining of Massive Datasets‖, Cambridge University Press, 2012.
4. Jeffrey D. Camm, James J. Cochran, Michael J. Fry, Jeffrey W. Ohlmann, David R. Anderson,―Essentials of Business Analytics‖, Cengage Learning, second Edition, 2016
5. U. Dinesh Kumar, ―Business Analytics: The Science of Data-Driven Decision Making‖, Wiley, 2017.

CA870402― Deep Learning

Instructional hours /week : 5

Total instructional hours   : 90

Credits   : 4

Module 1: **Introduction to Learning**

Limits of Traditional Computing – Machine Learning & Neural Networks – Neuron – FF Neural Networks – Types of Neurons – Softmax output layers

Module 2: **Deep Learning Models**

Tensor flow – Variables – Operations – Placeholders – Sessions – Sharing Variables – Graphs – Visualization

Module 3: **Convolutional Neural Network**

CNN Introduction – Feature Selection – Max Pooling – Filters and Feature Maps –Convolution Layer –Applications

Module 4: **Recurrent Neural Network**

RNN Introduction – Memory cells – sequence analysis – word2vec- LSTM — Memory

augmented Neural Networks – NTM—Application

Module 5: **Reinforcement Learning**

Reinforcement Learning – MDP – Q Learning – Applications

Reference Text

1. Nikhil Buduma, Nicholas Locascio, ―Fundamentals of Deep Learning: Designing NextGeneration Machine Intelligence Algorithms‖, O'ReillyMedia, 2017.
2. Ian Goodfellow, YoshuaBengio, Aaron Courville, ‖Deep Learning (Adaptive computation and Machine Learning series‖, MITPress, 2017.
3. Building Machine Learning Systems with Python, Richert & Coelho

CA030402 Project

Instructional hours /week : 10

Total instructional hours   : 180

Credits   : 5

Project aims at giving students hands-on experience in applying the programming knowledge in any language they have studied during this course to develop a real application / model for data analytics. Students must take up individual projects. Evaluation of the project is external.

CA030403 Comprehensive Viva - Voce

Credits   : 2

Course viva is conducted to find the skills the student has achieved by taking this programme.