Savitribai Phule Pune University

(Formerly University of Pune)

Department of Technology

STRUCTURE OF THREE YEAR FULL TIME DEGREE PROGRAM IN

B.Sc. Data Science

(Session 2022-23)

Semester	Lectures per Week	Course Credits
1st Semester	23	20
2nd Semester	23	20
3rd Semester	25	19
4th Semester	25	19
5th Semester	24	20
6th Semester	22	22
Total		Course Credits - 120

INDEX

In Seminar-1, Seminar-2 we will survey the fundamentals of data science by reading state of the art research papers in this area. This class will cover the basics of how to manipulate, integrate, and analyze data at scale. To receive credit, students must give in-class presentations and complete a final project.

B.Sc. Data Science

Semester-I

Subject Code	Subjects Name	Contact Hours per Week			Credits			
		L	P	T	L	P	T	Total
BDS101	Introduction to Data Science	3			3			3
BDS102	Applied Mathematics	3		1	3		1	4
BDS103	Data science with R	3			3			3
BDS104	Operating System and Information Security	3			3			3
BDS105	Data science with R Programming Lab		2X3=6			4		4
BDS106	English Communication-I	3			3			3
Total		15	6	2	15	4	1	20
Total Contact Hours per Week=23							**Total Credits=20**	

Semester-II

Subject Code	Subjects Name	Contact Hours per Week			Credits			
		L	P	T	L	P	T	Total
BDS201	Probability and Statistics	3		1	3		1	4
BDS202	Python Programming	3			3			3
BDS203	Data Science with Python	3			3			3
BDS204	Data Analytic and Visualization	3			3			3
BDS205	Python Programming LAB		2X3=6			4		4
BDS206	English Communication-II	3			3			3
Total		15	6	2	15	4	1	20
Total Contact Hours per Week=23							**Total Credits=20**	

B.Sc. Data Science
Semester-III

Subject Code	Subjects Name	Contact Hours per Week			Credits			
		L	P	T	L	P	T	Total
BDS301	Optimization Techniques	3			3			3
BDS302	Database Management System(DBMS)	3			3			3
BDS303	Machine Learning	3			3			3
BDS304	Optimization Techniques(Python)-LAB		2X2=4			2		2
BDS305	DBMS - LAB		2X3=6			3		3
BDS306	Machine Learning - Lab		2X3=6			3		3
BDS307	Seminar-1							2
Total		**9**	**16**	**0**	**9**	**8**		**19**
Total Contact Hours per Week=25								**Total Credits=19**

Semester-IV

Subject Code	Subjects Name	Contact Hours per Week			Credits			
		L	P	T	L	P	T	Total
BDS401	DeepLearning	3			3			3
BDS402	IoT Programming and Big Data	3			3			3
BDS403	Data warehouse and Data Mining	3			3			3
BDS404	Deep Learning - LAB		2X2=4			2		2
BDS405	IoT Programming and Big DataLAB		2X3=6			3		3
BDS406	Data warehouse and Data Mining LAB		2X3=6			3		3
BDS407	Seminar-2							2
Total		**9**	**16**	**0**	**9**	**8**		**19**
Total Contact Hours per Week=25								**Total Credits=19**

B.Sc. Data Science
Semester-V

Subject Code	Subjects Name	Contact Hours per Week			Credits			
		L	P	T	L	P	T	Total
BDS501	Big Data Analytics through Spark	3			3			3
BDS502	Introduction to Artificial Intelligence	3			3			3
BDS503	Machine Learning Operations (ML Ops)	3			3			3
BDS504	Elective – I	3			3			3
BDS505	Big Data Analytics through Spark-LAB		2X2=4			2		2
BDS506	Artificial Intelligence (PROLOG / Python) -LAB		2X2=4			2		2
BDS507	Project Work –Minor (IoT/Machine Learning)		2X2=4			4		4
Total		**12**	**12**		**12**	**8**		**20**
Total Contact Hours per Week=24							**Total Credits=20**	

Semester-VI

Subject Code	Subjects Name	Contact Hours per Week			Credits			
		L	P		L	P		Total
BDS601	NoSQL Databases	3			3			3
BDS602	Cloud Computing	3			3			3
BDS603	Big Data Acquisition And Analysis	3			3			3
BDS604	Elective - II	3			3			3
BDS605	Big Data Acquisition and Analysis Lab		2X2=4			2		2
BDS606	NoSQL Databases-LAB		2X2=4			2		2
BDS607	Project Work – Major		2X6=12			6		6
Total		**12**	**20**		**12**	**10**		**22**
Total Contact Hours per Week=22							**Total Credits=22**	

List of Electives

Elective-I

1. Technologies for Data Science
2. Computer Vision
3. Natural Language Processing and Text Mining

Elective-II

1. Health Analytics
2. Time Series Analysis and Forecasting
3. Product Development

B.Sc. Data Science
Semester-I

BDS101- Introduction to Data Science

Total Marks : 100		**Max. Time : 3 Hrs.**
End Sem Exam: 60 Marks	**Internal Assessment: 40 Marks**	**Credits:- 3**
Min. Pass Marks : 50%	**Min. Pass Marks : 50%**	**Lectures:- 45**

Learning Objectives: To make students able to

- Learn the fundamental concepts of data science
- Know the various domain and vertices of data science
- Learn the usage and application of data science

Learning Outcomes: After completion of this course successfully the students will be able to:

1. Understand the key difference between various areas of data science.
2. Understand the fundamental concepts of tool and techniques available in data science.
3. Understand the fundamental algorithms available in Artificial Intelligence.
4. Understand the key algorithms available in data mining and machine learning.

SECTION-A (15L)

Introduction to Data Science (15 L)

Foundation of Data science**,** Evolution of Data Science – Data Science Roles – Stages in a Data Science Project – Applications of Data Science in various fields – Data Security Issues.

Area and Scope of Data Science, Steps of Data Science Process: Data collection, Pre- processing, Data Collection Strategies – Data Pre-Processing Overview – Data Cleaning – Data Integration and Transformation – Data Reduction – Data Discretization. Training and testing. Use cases in various domain such Image, Natural Language, Audio and Video.

SECTION-B (15L)

Introduction to Artificial Intelligence:
Introduction Artificial Intelligence, The Foundations of AI, AI Technique, Production system characteristics, Production systems: 8-puzzle problem. Searching: Uniformed search strategies – Breadth first search, depth firstsearch.

SECTION-C (15L)

Searching Algorithms and Learning:
Local Search Algorithms: Generate and Test, Hill climbing, simulated annealing search, Constraint satisfactionproblems, Greedy best first search, A* search, AO* search.
Self-Learning: Propositional logic - syntax & semantics
Game Playing: Overview, Minimax algorithm, Alpha-Beta pruning, Additional Refinements.

Model Development
Simple and Multiple Regression – Model Evaluation using Visualization – Residual Plot – Distribution Plot – Polynomial Regression and Pipelines – Measures for In-sample Evaluation – Prediction and Decision Making.

SECTION-D (15L)

Introduction to Data Mining and Machine Learning: Introduction to Data Mining and Machine Learning, Supervised, Unsupervised and Reinforcement learning. Prediction vs Classification v/s Clustering. Association Rule Mining, classification and regression techniques,clustering, Scalability and data management issues in data mining algorithms, measures of interestingness.

Model Evaluation

Generalization Error – Out-of-Sample Evaluation Metrics – Cross Validation – Overfitting – Under Fitting and Model Selection – Prediction by using Ridge Regression – Testing Multiple Parameters by using Grid Search.

Recommended Books:

1. Rachel Schutt, Cathy O'Neil, "Doing Data Science: Straight Talk from the Frontiline" by Schroff/O'Reilly, 2013.
2. S. Russell and P. Norvig, Artificial Intelligence A Modern Approach, 2nd Edition. Pearson Education, 2007.
3. John W. Foreman, "Data Smart: Using data Science to Transform Information into Insight" by John Wiley &Sons, 2013.
4. Ian Ayres, "Super Crunchers: Why Thinking-by-Numbers Is the New Way to Be Smart" Ist Edition by Bantam,2007.
5. Eric Seigel, "Predictive Analytics: The Power to Predict who Will Click, Buy, Lie, or Die", 1st Edition, by Wiley,2013.
6. Matthew A. Russel, "Mining the Social Web: Data mining Facebook, Twitter, Linkedln,Goole+,
7. Ian Witten, Eibe Frank, Chris Pal and Mark Hall Data Mining: Practical Machine Learning Tools andTechniques.
8. 1. Jojo Moolayil, “Smarter Decisions : The Intersection of IoT and Data Science”, PACKT, 2016.
9. 2. Cathy O’Neil and Rachel Schutt , “Doing Data Science”, O'Reilly, 2015.
10. David Dietrich, Barry Heller, Beibei Yang, “Data Science and Big data Analytics”, EMC 2013
11. Raj, Pethuru, “Handbook of Research on Cloud Infrastructures for Big Data Analytics”, IGI Global

BDS102- Applied Mathematics

Total Marks : 100		**Max. Time : 3 Hrs.**
End Sem Exam: 60 Marks	**Internal Assessment: 40 Marks**	**Credits:- 3**
Min. Pass Marks : 50%	**Min. Pass Marks : 50%**	**Lectures:- 45**

Learning Objectives:

The course is a brief overview of the basic tools from Linear Algebra and Multivariable Calculus that will be needed in subsequent course of the program.

Learning Outcomes:

By completing the course the students will have been reminded of the basic tools of Linear Algebra and Multivariable Calculus needed in subsequent courses in the program notably:

- Fundamental properties of matrices, their norms, and their applications.
- Differentiating/Integrating multiple variable functions and the role of the gradient and the hessian matrix.
- Basic properties of optimization problems involving matrices and functions of multiple variables.

SECTION-A (5+5=10L)

1. Matrices and Basic Operations, Special structures Matrices and Basic Operations, Interpretation of matrices as linear mappings and some examples. Square Matrices, Determinants, Properties of determinants, singular and non-singular matrices, examples, finding an inverse matrix.

2. Introduction: Sets - finite and Infinite sets, Infinite Sets; functions, relations, Properties of Binary

3. Relations, Closure, Partial Ordering Relations; counting - Pigeonhole Principle, Permutation and Combination;Mathematical Induction, Principle of Inclusion and Exclusion.

4. Matrix and Basic properties of matrix & vectors:

Matrix, scalar multiplication, linear transformation, transpose, conjugate, rank, determinant, Inner and outer products, matrix multiplication rule and various algorithms, matrix inverse, square matrix, identitymatrix, triangular matrix, idea about sparse and dense matrix, unit vectors, symmetric matrix, Hermitian, skew-Hermitian and unitary matrices.

SECTION-B (8=8L)

5. Eigen values and Eigenvectors Characteristic Polynomial, Definition of Left/Right Eigen values and Eigenvectors, Caley – Hamilton theorem, singular value Decomposition, Interpretation of Eigen values/vectors.

Growth of Functions:

Asymptotic Notations, Summation formulas and properties, Bounding Summations, approximation by Integrals

Recurrences: Recurrence Relations, generating functions, Linear Recurrence Relations with constant coefficientsand their solution, Substitution Method, Recurrence Trees, Master Theorem

SECTION-C (8L)

6. Linear Systems Definition, applications, solving linear systems, linear inequalities, linear programming.

Graph Theory: Basic Terminology, Models and Types, multigraphs and weighted graphs, Graph Representation, Graph Isomorphism, Connectivity, Euler and Hamiltonian Paths and Circuits, Planar Graphs, Graph Coloring, Trees,Basic Terminology and properties of Trees, Introduction to Spanning Trees

Linear Transformations

Definition and example of linear transformation, Null space, range, rank and nullity of linear transformation,matrix representation of a linear transformation, dual space, dual basis, double dual, composition of linear transformation and matrix multiplication.

Transformation Diagonalization: Diagonalizability, matrix Limits and Markov Chains and the Caley-Hamilton Theorem.

SECTION-D (5+8=18L)

7. Real-valued functions of two or more variables. Definition, examples, simple demos, applications.
Prepositional Logic: Logical Connectives, Well-formed Formulas, Tautologies, Equivalences, Inference Theory

8. Analysis elements Distance, Limits, Continuity, Differentiability, the gradient and the Gaussian. Optimization problems Simple examples, motivation, the role of the Hessian maxima and minima and related extreme conditions. Integration Double integrals, Fubini's theorem, properties, applications.
9. Numerical Linear Algebra: Regularization, Principal Component Analysis, Singular-Value, Decomposition, Latent Semantic Analysis,
Case Studies: Recommender Systems, Page Ranking

Recommended Books:

1. Gilbert Strang, Linear Algebra and its Applications. Thomson /Brooks Cole (Available in a Greek Translation).
2. Thomas M. Apostol, Calculus, Wiley, 2nd Edition, 1991 ISBN 960-07-0067-2.
3. Michael Spivak. Calculus, publish or Perish, 2008, ISBN 978-0914098911.
4. Ross L. Finney, Maurice D.Weir . and Frank R. Giordano. Thomas's Calculus, Pearson 12th Edition 2009.
5. David C. Lay, Linear Algebra and Its Applications, 4th Editoin.
6. Yourself saad, Iterative Methods for spare Linear Systems.
7. C.L. Liu, D.P. Mahopatra, Elements of Discrete mathematics, 2nd Edition, Tata McGraw Hill, 1985.
8. Kenneth Rosen, Discrete Mathematics and Its Applications, Sixth Edition, McGraw Hill 2006
9. T.H.Coremen, C.E.Leiserson, R. L. Rivest, Introduction to algorithms, 3rd edition Prentice Hall on India,
10. M. O. Albertson and J. P. Hutchinson, Discrete Mathematics with Algorithms, John wiley Publication, 1988
11. J. L. Hein, Discrete Structures, Logic, and Computability, 3rd Edition, Jones and Bartlett Publishers, 2009
12. D.J. Hunter, Essentials of Discrete Mathematics, Jones and Bartlett Publishers, 2008
13. Stephen H. Friedberg, Arnold J. Insel, Lawrence E. Spence, Linear Algebra, 4th Ed., PrenticeHall of India Pvt. Ltd., New Delhi, 2004.
14. S. Lang, Introduction to Linear Algebra, 2nd Ed., Springer, 2005.
15. A.I. Kostrikin, Introduction to Algebra, Springer Verlag, 1984.
16. Richard Bronson, Theory and Problems of Matrix Operations, Tata McGraw Hill, 1989.

Applied Mathematics Tutorial

Student Activity:

1. Find the Eigenvectors of A = { 1 1 1 1,2 3 4 5,3 4 5 6}
2. Find orthogonal S =Spam {(1 1 1 1),(1 4 4 0),(-1 4 4 0),(-4 2 2 0)}

Tutorial:

1. Study various applications of Matrices.
2. Study different polynomial functions and their uses.
3. Take one real world example and apply the Linear System solution.
4. Study some real valued functions and its applications.
5. Study and solve one optimization problem.

BDS103 - DATA SCIENCE WITH R

Total Marks :100		**Max. Time : 3 Hrs.**
End Sem Exam : 60 Marks	**Internal Assessment : 40 Marks**	**Credits:- 3**
Min. Pass Marks : 50%	**Min. Pass Marks : 50%**	**Lectures:- 45**

Learning Objective:
Data Science is a fast-growing interdisciplinary field, focusing on the analysis of data to extract knowledge and insight. This course will introduce students to the collection. Preparation, analysis, modeling and visualization of data, covering both conceptual and practical issues. Examples and case studies from diverse fields will be presented, and hands-on use of statistical and data manipulation software will be included.

Learning Outcomes:
1. Recognize various disciplines that contribute to a successful data science effort.
2. Understand the processes of data science - identifying the problem to be solved, data collection, preparation, modeling, evaluation and visualization.
3. Be aware of the challenges that arise in data sciences.
4. Develop and appreciate various techniques for data modeling and mining.
5. Be cognizant of ethical issues in many data science tasks.

Unit-1: Basics of R-Programming

Evolution of R, Features of R, Local Environment support, R Command prompt, R Script File, Comment, R Data types, R Variables, R Operators-function.
Understanding data: Introduction – Types of Data: Numeric – Categorical – Graphical – High Dimensional Data – Classification of digital Data: Structured, Semi-Structured and Unstructured - Example Applications. Sources of Data: Time Series – Transactional Data – Biological Data – Spatial Data – Social Network Data – Data Evolution.

Unit-2: R Fundamentals

Introduction to R- Features of R - Environment - R Studio. R-Decision Making:- R-If statement, R-If….else statement, R- The if….else if…else statement-Switch Statement, R-Loop:- Repeat loop, While loop, for loop, Loop ,Control statement:- Break, Next.
Basics of R-Assignment - Modes -Operators - special numbers - Logical values – Basic Functions - R help functions - R Data Structures - Control Structures.
Vectors: Definition- Declaration - Generating - Indexing -Naming - Adding & Removing elements - Operations on Vectors - Recycling - Special Operators - Vectorized if- then else-Vector Equality – Functions for vectors - Missing values - NULL values - Filtering & Subsetting.

Unit-3: Data Structures in R

Matrices - Creating Matrices - Adding or Removing rows/columns - Reshaping - Operations - Special functions on Matrices. Lists - Creating List – General List Operations - Special Functions - Recursive Lists. Data Frames - Creating Data Frames - Naming - Accessing - Adding - Removing - Applying Special functions to Data Frames - Merging Data Frames Factors and Tables.

Unit- 4: Working With Data in R

Input / Output – Reading and Writing datasets in various formats - Functions - Creating User defined functions - Functions on Function Object - Scope of Variables - Accessing Global, Environment - Closures - Recursion. Exploratory Data Analysis - Data Preprocessing -

Descriptive Statistics - Central Tendency - Variability - Mean - Median - Range - Variance - Summary - Handling Missing values and Outliers - Normalization
Data Visualization in R : Types of visualizations - packages for visualizations - Basic Visualizations, Advanced Visualizations and Creating 3D plots.

Unit- 5: Statistics in R

Inferential Statistics with R - Types of Learning - Linear Regression- Simple Linear Regression - Implementation in R - functions on lm() - predict() - plotting and fitting regression line. Multiple Linear Regression - Introduction -comparison with simple linear regression - Correlation Matrix - F-Statistic - Target variables Vs Predictors - Identification of significant features - Implementation of Multiple Linear Regression in R. R-Reshaping: - Joining rows and columns, merging data frames, melting and casting. R- CSV Files: - Getting and starting with directory, Input as a CSV file, Reading CSV file, Analyzing CSV file, writing to CSV file. R- EXCEL File:- Install xlsx Packages, Verify & Load "xlsx" packages, Input as a xlsx file, Reading excel file. R-Binary File:- writing binary file, reading binary file. R- XML File:- Input data, Reading XML file, details of the first node, xml to data node.

Unit 6:- Application of R- programming

R- Pie charts: - Pie chart title and colour, 3-D Pie Chart. R- Bar Chart: - Bar Chart Labels, Title and colour, Group Bar chart and stacked bar chart. R- Box Plot: - Creating a box plot, Box plot with notch. R- Histogram: - Range of x and y values.

References

1. Nina Zumel, John Mount, "Practical Data Science with R", Manning Publications, 2014.
2. Jure Leskovec, Anand Rajaraman, Jeffrey D.Ullman,"Mining of Massive Datasets", Cambridge University Press, 2014.
3. Mark Gardener, "Beginning R - The Statistical Programming Language", John Wiley & Sons, Inc., 2012.
4. W. N. Venables, D. M. Smith and the R Core Team, "An Introduction to R", 2013.
5. Tony Ojeda, Sean Patrick Murphy, Benjamin Bengfort, Abhijit Dasgupta,"Practical Data Science Cookbook", Packt Publishing Ltd., 2014.
6. Nathan Yau, "Visualize This: The FlowingData Guide to Design, Visualization, and Statistics", Wiley, 2011.
7. Boris lublinsky, Kevin t. Smith, Alexey Yakubovich, "Professional Hadoop Solutions", Wiley, ISBN: 9788126551071, 2015.

Student Activity

Databases need to undergo pre-processing to be useful for data mining. Dirty data can cause confusion for the data mining procedure, resulting in unreliable output. Data cleaning includes smoothing noisy data, filling in missing values, identifying and removing outliers, and resolving inconsistencies.

RECOMMENDED CO-CURRICULAR ACTIVITIES:

(Co-curricular activities shall not promote copying from textbook or from others work and shall encourage self/independent and group learning)

A. Measurable

1. Assignments (in writing and doing forms on the aspects of syllabus content and outside the syllabus content. Shall be individual and challenging)
2. Student seminars (on topics of the syllabus and related aspects (individual activity))
3. Quiz (on topics where the content can be compiled by smaller aspects and data (Individuals or

groups as teams))
4. Study projects (by very small groups of students on selected local real-time problems pertaining to syllabus or related areas. The individual participation and contribution of students shall be ensured (team activity

B. General

1. Group Discussion
2. Try to solve MCQ's available online.
3. Others

RECOMMENDED CONTINUOUS ASSESSMENT METHODS:

Some of the following suggested assessment methodologies could be adopted;

1. The oral and written examinations (Scheduled and surprise tests)
2. Closed-book and open-book tests
3. Problem-solving exercises
4. Practical assignments and laboratory reports
5. Observation of practical skills
6. Individual and group project reports like "COVID-19 Analysis", "Estimated Quarantine Period for Covid-19 Contacts", etc.
7. Efficient delivery using seminar presentations,
8. Viva voce interviews.
9. Computerized adaptive testing, literature surveys and evaluations,
10. Peers and self-assessment, outputs form individual and collaborative work

BDS104 - Operating System and Information Security

Total Marks : 100		Max. Time : 3 Hrs.
End Sem Exam: 60 Marks	Internal Assessment: 40 Marks	Credits:- 3
Min. Pass Marks : 50%	Min. Pass Marks : 50%	Lectures:-45

Learning Objectives:

1. To understand the fundamental concepts and techniques of Operating Systems.
2. To study the concepts in process management and concurrency control mechanisms.
3. To understand the concepts in memory managements and deadlocks solutions.
4. To study on file management and storage structures

Learning Outcomes: On successful completion of this course, the students are able:

1. To understand basic concepts of operating system.
2. To describe process management, scheduling and concurrency control mechanisms.
3. To analyze memory management and deadlocks.
4. To compare various file systems with operating systems examples.

SECTION-A (10L)

1. Introduction:

Basic OS functions, resource abstraction, types of operating systems–programming systems, batch systems, timesharing systems; operating systems for personal computers & workstations, process control & real time systems.

2. Operating System Organization (6 Lectures) Processor and user modes, kernels, system calls and system courses.

3.Process Management: System view of the process and resources, process abstraction, process hierarchy, threads, threading issues, thread libraries; Process Scheduling, non-pre-emptive and pre-emptive scheduling algorithms; concurrent and processes, critical section, semaphores, methods for interprocess communication; deadlocks.

SECTION-B (20L)

1.Memory Management: Physical and virtual address space; memory allocation strategies –fixed and variable partitions, paging, segmentation, virtual memory.

2.File and I/O Management: Directory structure, file operations, files allocation methods, device management.

3. Protection and Security: Policy mechanism, Authentication, Internal access Authorization.

SECTION-C (10L)

The Security Problem in Computing: The meaning of computer Security, Computer Criminals, Methods of Defense, Elementary Cryptography: Substitution Ciphers, Transpositions, Making "Good" Encryption algorithms, The Data Encryption Standard, The AES Encryption Algorithms, Public Key Encryptions, Uses of Encryption. Program Security: Secure Programs, Nonmalicious Program Errors, viruses and other malicious code, Targeted Malicious code, controls Against Program Threats, Protection in General- Purpose operating system protected objects and methods of protection memory and addmens protection, File protection Mechanisms, User Authentication Designing Trusted O.S: Security polices, models of security, trusted O.S design, Assurance in trusted O.S. Implementation examples

SECTION-D (10+4=14L)

Data base Security: Security requirements, Reliability and integrity, Sensitive data, Inference, multilevel database, proposals for multilevel security. Security in Network: Threats in Network, Network Security Controls, Firewalls, Intrusion Detection Systems,Secure E-Mail.

Administering Security: Security Planning, Risk Analysis, Organizational Security policies, Physical Security.

Legal Privacy and Ethical Issues in Computer Security: Protecting Programs and data, Information and the law, Rights of Employees and Employers, Software failures, Computer Crime, Praia, Ethical issues in Computer Security, case studies of Ethics.

Recommended Books:

1. A Silberschatz, P.B. Galvin, G. Gagne, Operating Systems Concepts, 8th Edition, John Wiley Publications 2008.
2. A.S. Tanenbaum, Modern Operating Systems, 3rd Edition, Pearson Education 2007.
3. G. Nutt, Operating Systems: A Modern Perspective, 2nd Edition Pearson Education 1997.
4. W. Stallings, Operating Systems, Internals & Design Principles, 5th Edition, Prentice Hall of India. 2008.
5. M. Milenkovic, Operating Systems- Concepts and design, Tata McGraw Hill 1992.

BDS105- R PROGRAMMING LAB

Total Marks : 50		Max. Time : 2 Hrs.
End Sem Exam : 30 Marks	Internal Assessment : 20 Marks	Credits:- 2
Min. Pass Marks : 50%	Min. Pass Marks : 50%	Lectures:- 30

R Programming LAB

1) Installing R and R studio
2) Create a folder DS_R and make it a working directory. Display the current working directory
3) Installing the "ggplot2", "caTools", "CART" packages
4) Load the packages "ggplot2", "caTools".
5) Basic operations in r
6) Working with Vectors:
   - Create a vector v1 with elements 1 to 20.
   - Add 2 to every element of the vector v1.
   - Divide every element in v1 by 5
   - Create a vector v2 with elements from 21 to 30. Now add v1 to v2.

7) Getting data into R, Basic data manipulation
8) Using the data present in the table given below, create a Matrix "M"

	C1	C2	C3	C4	C5
C1	0	12	13	8	20
C2	12	0	15	28	88
C3	13	15	0	6	9
C4	8	28	6	0	33
C5	20	88	9	33	0

   - Find the pairs of cities with shortest distance.

9) Consider the following marks scored by the 6 students

Section	Student no	M1	M2	M3
A	1	45	54	45
A	2	34	55	55
A	3	56	66	64
B	1	43	44	45
B	2	67	76	78
B	3	76	68	37

- create a data structure for the above data and store in proper positions with proper names
- display the marks and totals for all students
- Display the highest total marks in each section.
- Add a new subject and fill it with marks for 2 sections.
- Three people denoted by P1, P2, P3 intend to buy some rolls, buns, cakes and bread. Each of them needs these commodities in differing amounts and can buy them in two shops S1, S2. The individual prices and desired quantities of the commodities are given in the following table "demand.

	price	
	S1	S2
Roll	1.5	1
Bun	2	2.5
Cake	5	4.5
Bread	16	17

	demand.quantity			
	Roll	Bun	Cake	Bread
P1	6	5	3	1
P2	3	6	2	2
P3	3	4	3	1

- Create matrices for above information with row names and col names.
- Display the demand.quantity and price matrices
- Find the total amount to be spent by each person for their requirements in each shop
- Suggest a shop for each person to buy the products which is minimal.

10) Consider the following employee details:

employee details as follows		
	emp_no:1	
	name: Ram	
	salary	
		basic: 10000
		hra: 2500
		da: 4000
	deductions	
		pf: 1100
		tax: 200
	total salary	
		gs(Gross Salary):
		ns(Net Salary)

- Create a list for the employee data and fill gross and net salary.
- Add the address to the above list
- display the employee name and address
- remove street from address
- remove address from the List.

11) Loops and functions - Find the factorial of a given number
12) Implementation of Data Frame and its corresponding operators and functions
13) Implementation of Reading data from the files and writing output back to the specified file
14) Treatment of NAs, outliers, Scaling the data, etc
15) Applying summary() to find the mean, median, standard deviation, etc
16) Implementation of Visualizations - Bar, Histogram, Box, Line, scatter plot, etc.
17) Implementation of Linear and multiple Linear Regression
18) Fitting regression line

BDS106 - English Communications/Technical Communications

Total Marks : 50		**Max. Time : 2 Hrs.**
End Sem Exam : 30 Marks **Min. Pass Marks : 50%**	**Internal Assessment : 20 Marks** **Min. Pass Marks : 50%**	**Credits:- 2** **Lectures:- 30**

Learning Objectives:

The purpose of this course is twofold: to train students in communication skills and to help develop in them a facility for communicative English. Since language it is which binds society together and serves as a crucial medium of interaction as wellas interchange of ideas and thoughts, it is important that students develop a capacity for clear and effective communication, spoken and written.

Learning Outcomes: On completion of this course students should be able to:

To unlock the communicator in them by using English appropriately and with confidence for further studies or in professional spheres where English is the indispensable tool of communication.

Unit 1 [10] Introduction

1. What is communication?
2. Types of communication: Horizontal, Vertical, Interpersonal, Grapevine
3. Uses of Communication

Prescribed Reading: Chapter 1 Applying Communication Theory for Professional Life: A Practical Introductionby Dainton and Zelley

http://tsime.uz.ac.zw/claroline/backends/download.php?url=L0ludHJvX3RvX2NvbW11bmljYXRpb25fVGhlb3J5LnBkZg%3 D%3D&cidReset=true&cidReq=MBA563

Unit 2 [10] Language of Communication

1. Verbal: spoken and written
2. Non-verbal: Proxemics, Kinesics, Haptics, Chronemics, Paralinguistic
3. Barriers to communication

Unit 3 [10] Reading Comprehension

1. Locate and remember the most important points in the reading
2. Interpret and evaluate events, ideas, and information
3. Read “between the lines” to understand underlying meanings
4. Connect information to what they already know

Unit 4 [10] Writing

1. Expanding an Idea
2. Note Making
3. Memo
4. Writing Formal Email
5. Writing a Business Letter
6. Report Writing

B.Sc. Data Science
Semester-II
BDS201 - Probability and Statistics

Total Marks : 100		**Max. Time : 3 Hrs.**
End Sem Exam: 60 Marks	**Internal Assessment: 40 Marks**	**Credits:- 3**
Min. Pass Marks : 50%	**Min. Pass Marks : 50%**	**Lectures:- 45**

Learning Objectives: To make students able to

1. Learn the basics and advance concepts of Probability theory
2. Learn various sampling techniques
3. Find and understand the applications of Probabilities in data science

Learning Outcomes: After completion of this course successfully the students will be able to:

1. Solve problem related to probability
2. Understand and use various Probabilities theory to solve real problem of data science
3. Understand and use various Probability distributions for different machine learning related task
4. Understand the sampling techniques and use them for preparing effective datasets.

SECTION-A (15L)

Basic Probability - Random Experiments - Sample Spaces Events - The Concept of Probability -The Axioms of Probability -

Some Important Theorems on Probability - Assignment of Probabilities -Conditional Probability -Theorems on Conditional Probability – Independent Events -Bayes' Theorem or Rule Combinatorial Analysis - Fundamental Principle of Counting - Tree Diagrams –Permutations.

Introduction to Statistics – Primary and Secondary data – Nominal, Ordinal, Ratio, and Interval scale (with examples) - Graphical Representation of data – Bar-charts, Pie-diagrams, Histograms, Frequency polygon, Ogives**.**

Central Limit Theorem and Confidence Interval:

Introduction, Sampling Variability and CLT, CLT (for the mean) examples, Confidence Interval (for a mean), Accuracy vs. Precision, Required Sample Size for ME, CI (for the mean) examples.

SECTION-B (15L)

Random Variables and Probability Distributions - Random Variables - Discrete Probability Distributions -Distribution Functions for Random Variables - Distribution Functions for Discrete Random Variables - Continuous Random Variables – Graphical Interpretations Joint Distributions Independent Random Variables - Change of Variables - Probability Distributions of Functions of Random Variables – Convolutions – Conditional Distributions Applications to Geometric Probability.

Measures of central tendency :– properties – merits and demerits – weighted means– graphical location of median,quartiles, deciles, percentiles, and mode – relation between arithmetic mean, geometric mean and harmonic mean.

Measures of dispersion : – characteristics – Coefficient of dispersion – Coefficient of variation – Moments –Relation between moments about mean in terms of moments about point – Pearson's coefficients.

Inference and Significance: Introduction to Inference, Hypothesis Testing (for a mean), HT (for the mean) examples, Inference for Other Estimators, Decision Errors, Significance vs. Confidence Level, Statistical vs. Practical Significance.

SECTION-C (15L)

Mathematical Expectation - Definition of Mathematical Expectation - Functions of Random Variables - Theorems on Expectation - Variance & Standard Deviation - Theorems on Variance - Standardized Random Variables - Special Probability Distributions - Binomial Distribution - Normal Distribution - Poisson distribution.

Skewness and Kurtosis – Pearson's coefficient of skewness – Bowley's coefficient of skewness – coefficient ofskewness based upon moments

Curve fitting – Principle of least squares – Fitting of straight line, parabola, exponential and power curve.
Inference for Comparing Means: Introduction, t-distribution, Inference for a mean, Inference for comparing two independent means, Inference for comparing two paired means, Power, Comparing more than two means, ANOVA, Conditions for ANOVA, Multiple comparisons, Bootstrapping.

SECTION-D (15L)

Sampling Theory - Population and Sample - Statistical Inference- Sampling With and Without Replacement Random Samples - Random Numbers - Population Parameters - Sample Statistics -Sampling Distributions - Sample Mean - Sampling Distribution of Means - Sampling Distribution of Proportions - Sampling Distribution of Differences and Sums – Sample Variance - Sampling Distribution of Variances - Computation of Mean, Variance, and Moments for Grouped Data - The Least-Squares Parabola - Multiple Regression Standard Error of Estimate The Linear Correlation Coefficient Generalized CorrelationCoefficient Rank Correlation
Inference for Proportions: Introduction, Sampling Variability and CLT for Proportions, Confidence Interval for a Proportion, Hypothesis Test for a Proportion, Estimating the Difference Between Two Proportions, Hypothesis Test for Comparing Two Proportions, Small Sample Proportions, Examples, Comparing Two Small Sample Proportions, Chi-Square GOF Test, The Chi-Square Independence Test.

Correlation and Regression:
Simple correlation – Karl Pearson's coefficient. of correlation – Rank correlation –Simple Regression – lines ofregression – properties of regression coefficient –Multiple and Partial correlation coefficient in three variables. **Hypothesis Testing :**
Estimation and Hypothesis testing, t-test, chi-square test, ANOVA

Recommended Books:

1. Murray R. Spiegel, John J. Schiller & R. Alu Srinivasan, "Probability and Statistics", Schaum outlines, McGraw Hill, 3rd edition, 2009.
2. S. P. Gupta, Statistical Methods, S. Chand and Sons.
3. S. C Gupta and V. K. Kapoor, "Fundamentals of Mathematical Statistics", 11th edition, S.Chand and Sons.
4. Agarwal.B.L (1996): Basic Statistics, 3/e, New Age International (P) Ltd.,.
5. Sanjay Arora & Bansilal (2002): New Mathematical statistics, Meerat Publications, New Delhi
6. Hooda.R.P.(2003): Statistics for Business and Economics, 3/e, Mac Millan.

BDS202 – Python Programming

Total Marks : 100		**Max. Time : 3 Hrs.**
End Sem Exam : 60 Marks	**Internal Assessment : 40 Marks**	**Credits:- 3**
Min. Pass Marks : 50%	**Min. Pass Marks : 50%**	**Lectures:- 45**

Learning Objectives:

1. To impart the basic concepts of data structures and algorithms.
2. To understand concepts about searching and sorting techniques
3. To Understand basic concepts about stacks, queues, lists, trees and graphs
4. To understand the algorithms and develop the step by step solutions of problems with the help of data structures.

Learning Outcomes: On successful completion of this course, the students are able:

1. To analyze algorithms and their correctness.
2. To implement various searching and sorting techniques in different problems.
3. To have knowledge of concepts related with tree and graphs and able to apply them.

Unit-1: Python Programming 10 Hrs

General Problem Solving Concepts- Problem solving in everyday life, types of problems, problem solving with computers, difficulties with problem solving, problem solving aspects, top down design. Problem Solving Strategies, Program Design Tools: Algorithms, Flowcharts and Pseudo-codes, implementation of algorithms. Basics of Python Programming: Features of Python, History and Future of Python, Writing and executing Python program, Literal constants, variables and identifiers, Data Types, Input operation, Comments, Reserved words, Indentation, Operators and expressions, Expressions in Python. Decision Control Statements: Decision control statements, Selection/conditional branching Statements: if, if-else, nested if, if-elif-else statements. Basic loop Structures/Iterative statements: while loop, for loop, selecting appropriate loop. Nested loops, the break, continue, pass, else statement used with loops. Other data types- Tuples, Lists and Dictionary. Need for functions, Function: definition, call, variable scope and lifetime, the return statement. Defining functions, Lambda or anonymous function, documentation string, good programming practices. Introduction to modules, Introduction to packages in Python, Introduction to standard library modules.

Unit-2: Strings Linear Data Structures 8 Hrs

Strings and Operations- concatenation, appending, multiplication and slicing. Strings are immutable, strings formatting operator, built in string methods and functions. Slice operation, ord() and chr() functions, in and not in operators, comparing strings, Iterating strings, the string module. **Concept** of Sequential Organization, Overview of Array, Array as an Abstract Data Type, Operations on Array, Merging of two arrays, Storage Representation and their Address Calculation: Row major and Column Major, Multidimensional Arrays: Two-dimensional arrays, n-dimensional arrays. Concept of Ordered List.

Unit-3: Object Oriented Programming & File Handling 8 Hrs

Programming Paradigms-monolithic, procedural, structured and object oriented, Features of Object oriented programming-classes, objects, methods and message passing, inheritance, Polymorphism, containership, reusability, delegation, data abstraction and encapsulation.

Classes and Objects: classes and objects, class method and self-object, class variables and

object variables, public and private members, class methods. File Handling and Dictionaries Files: Introduction, File path, Types of files, Opening and Closing files, Reading and Writing files. Dictionary method. Dictionaries- creating, assessing, adding and updating values. Case Study: Study design, features, and use of any recent, popular and efficient system developed using Python. (This topic is to be excluded for theory examination).

Unit-4: Linked List, Stacks 08 Hrs

Introduction to Static and Dynamic Memory Allocation, **Linked List:** Introduction, of Linked Lists, Realization of linked list using dynamic memory management, operations, Linked List as ADT, **Types of Linked List:** singly linked, linear and Circular Linked Lists, Doubly Linked List, Doubly Circular Linked List. Basic concept, stack Abstract Data Type, Representation of Stacks Using Sequential Organization, stack operations, Multiple Stacks, Applications of Stack- Expression Evaluation and Conversion, Polish notation and expression conversion.

Unit-5: Queue 07 Hrs

Basic concept, Queue as Abstract Data Type, Representation of Queue using Sequential organization, Queue Operations, Circular Queue and its advantages, Multi-queues, Linked Queue and Operations. Deque-Basic concept, types (Input restricted and Output restricted), Priority Queue- Basic concept, types (Ascending and Descending).

Unit-6: Working with Data in Python 07 Hrs

Introduction, Working with NumPy Arrays, examples of using NumPy array manipulation to access data and subarrays, and to split, reshape, and join the arrays. Creating matrices, Transposing and reshaping a matrix, Importing and exporting a CSV, Plotting arrays with Matplotlib.

References:

1. Python Data Science Handbook - Essential Tools for Working with Data (Jake VanderPlas)
2. Data Science And Analytics With Python (Jesus Rogel - Salazar)
3. Mastering Python for Data Science (Madhavan Samir
4. R. G. Dromey, “How to Solve it by Computer”, Pearson Education India; 1st edition, ISBN10: 8131705625, ISBN-13: 978-8131705629 Maureen Spankle, “Problem
5. Solving and Programming Concepts”, Pearson; 9th edition, ISBN-10: 9780132492645, ISBN-13: 978- 0132492645
6. Romano Fabrizio, “Learning Python”, Packt Publishing Limited, ISBN: 9781783551712, 1783551712
7. Paul Barry, “Head First Python- A Brain Friendly Guide”, SPD O’Reilly, 2nd Edition, ISBN:978-93-5213-482-3
8. Martin C. Brown, “Python: The Complete Reference”, McGraw Hill Education, ISBN-10: 9789387572942, ISBN-13: 978-9387572942, ASIN: 9387572943
9. Jeeva Jose, P. Sojan Lal, “Introduction to Computing & Problem Solving with Python”, Khanna Computer Book Store; First edition, ISBN-10: 9789382609810, ISBN-13: 978- 9382609810

Text Books:

1. Reema Thareja, “Python Programming Using Problem Solving Approach”, Oxford University Press, ISBN 13: 978-0-19-948017-6
2. R. Nageswara Rao, “Core Python Programming”, Dreamtech Press; Second edition ISBN10: 938605230X, ISBN-13: 978-9386052308 ASIN: B07BFSR3LL

BDS205 - Data Science with Python

Total Marks : 100		**Max. Time : 3 Hrs.**
End Sem Exam : 60 Marks	**Internal Assessment : 40 Marks**	**Credits:- 3**
Min. Pass Marks : 50%	**Min. Pass Marks : 50%**	**Lectures:- 45**

Learning Objectives:

5. To impart the basic concepts of data structures and algorithms.
6. To understand concepts about searching and sorting techniques
7. To Understand basic concepts about stacks, queues, lists, trees and graphs
8. To understand the algorithms and develop the step by step solutions of problems with the help of data structures.

Learning Outcomes: On successful completion of this course, the students are able:

1. To analyze algorithms and their correctness.
2. To implement various searching and sorting techniques in different problems.
3. To have knowledge of concepts related with tree and graphs and able to apply them.

SECTION-A (5+5+10=20L)

Data Structures and OOP Python Program Execution Procedure – Statements – Expressions – Flow of Controls – Functions – Numeric Data Types – Sequences – Strings – Tuples – Lists – Dictionaries. Class – Constructors – Object Creation – Inheritance – Overloading. Text Files and Binary Files – Reading and Writing.

SECTION-B (5+5=10L)

Numpy and Pandas Packages NumPy ndarray - Vectorization Operation - Array Indexing and Slicing - Transposing Array and Swapping Axes - Saving and Loading Array - Universal Functions - Mathematical and Statistical Functions in Numpy . Series and DataFrame data structures in pandas - Creation of Data Frames – Accessing the columns in a DataFrame - Accessing the rows in a DataFrame - Panda's Index Objects - Reindexing Series and DataFrames - Dropping entries from Series and Data Frames - Indexing, Selection and Filtering in Series and Data Frames - Arithmetic Operations between Data Frames and Series - Function Application and Mapping.

SECTION-C (20L)

Data Wrangling Combining and Merging Data Sets – Reshaping and Pivoting – Data Transformation – String Manipulations – Regular Expressions.

Data Aggregation and Group Operations Group By Mechanics – Data Aggregation – GroupWise Operations – Transformations – Pivot Tables – Cross Tabulations – Date and Time data types.

SECTION-D (5+5=10L)

Visualization in Python Matplotlib and Seaborn Packages – Plotting Graph - Controlling Graphs – Adding Text – More Graph Types – Getting and Setting Values – Patches.

REFERENCES:

1. Gowrishanker and Veena, "Introduction to Python Programming", CRC Press, 2019.
2. Python Crash Course, 2nd Edition, By Eric Matthes, May 2019
3. NumPy Essentials, By Leo Chin and Tanmay Dutta, April 2016
4. Joel Grus, "Data Science from scratch", O'Reilly, 2015.
5. Wes Mc Kinney, "Python for Data Analysis", O'Reilly Media, 2012.
6. Kenneth A. Lambert, (2011), "The Fundamentals of Python: First Programs", Cengage Learning
7. Jake Vanderplas. Python Data Science Handbook: Essential Tools for Working with Data 1st Edition.

BDS206- Data Analytic and Visualization

Total Marks : 100		**Max. Time : 3 Hrs.**
End Sem Exam: 60 Marks	**Internal Assessment: 40 Marks**	**Credits:- 3**
Min. Pass Marks : 50%	**Min. Pass Marks : 50%**	**Lectures:- 45**

Learning Objectives:

1. Make students understand about commonly used terms and techniques related to data analytics that are inuse today.
2. To discuss and train students on how data and data analytics can be used by managers to make betterdecisions.
3. Have the student gain perspective and practice by applying data analysis techniques in several settings.
4. Make students learn to use Python and its libraries to perform various Data Analytic tasks

Learning Outcomes: After completion of this course successfully the students will be able to:

1. Understand the need and importance of decision making in business, its inherent difficulties and pitfalls,
2. Learn the importance of proper data analysis in decision making.
3. Understand how the data environment in various domain is changing
4. Apply common quantitative and visual techniques to enhance decision making.
5. Use Python to analyze data and provide useful information for decision making.
6. Tableau course syllabus will help you to become master in Business Intelligence (BI) tool, Data Visualization, reporting and SQL with real-life industry Projects in Health care, Retail and Banking domains.Latest Tableau course content to pass Tableau Desktop, Analyst and Server certification exams.

SECTION-A (10L)

Data Definitions and Analysis Techniques: Elements, Variables, and Data categorization, Levels of Measurement, Data management and indexing,Introduction to statistical learning.
Basic analysis techniques: Statistical hypothesis generation and testing, Chi-Square test , t-Test , Analysis of variance , Correlation analysis ,Maximum likelihood test.
Data analysis techniques: Regression analysis , Classification techniques , Clustering , Association rules analysis

SECTION-B (10L)

Learn Tableau Basic Reports
Parameters, Grouping Example 1, Grouping Example 2, Edit Groups, Set, Combined Sets, Creating a First Report, Data Labels, Create Folders, Sorting Data, Add Totals, Sub Totals and Grand Totals to Report.
Learn Tableau Charts
Area Chart, Bar Chart, Box Plot, Bubble Chart, Bump Chart, Bullet Graph, Circle Views, Dual Combination Chart, Dual Lines Chart, Funnel Chart, Traditional Funnel Charts, Gantt Chart, Grouped Bar or Side by Side Bars Chart, Heatmap, Highlight Table, Histogram, Cumulative Histogram, Line Chart, Lollipop Chart, Pareto Chart, Pie Chart, Scatter Plot, Stacked Bar Chart, Text Label, Tree Map, Word Cloud, Waterfall Chart, Geographic map, Filled map, Crosstab, Combines axis, Motion chart, Reference lines

SECTION-C (10L)

SQL Convert to Custom SQL, **Learn Tableau Advanced Reports:** Dual Axis Reports, Blended Axis, Individual Axis, Add Reference Lines, Reference Bands, Reference Distributions, Basic Maps, Symbol Map, Use Google Maps, Mapbox Maps as a Background Map, WMS Server Map as a Background Map

Learn Tableau Calculations & Filters
Calculated Fields, Basic Approach to Calculate Rank, Advanced Approach to Calculate Rank, Calculating Running Total, Filters Introduction, Quick Filters, Filters on Dimensions, Conditional Filters, Top and Bottom Filters, Filters on Measures, Context Filters, Slicing Filters, Data Source Filters, Extract Filters.

SECTION-D (10L)

Learn Tableau Dashboards
Create a Dashboard, Format Dashboard Layout; Create a Device Preview of a Dashboard, Create Filters on Dashboard, Dashboard Objects, and Create a Story.
Tableau data server: Physical architecture overview, User access, Component functions & processes, Tableau server on-premises, Tableau reader, Tableau online v tableau server.
Tableau Server UI: Tableau server user interface, Users, Site roles, Groups, Schedules, Tasks, Tableau server menu, Content display options.

Recommended Books:

1. All of statistics: a concise course in statistical inference. Larry Wasserman. Springer, 2004.
2. C. Bishop, Pattern Recognition and Machine Learning, Springer 2007
3. Hastie, Trevor, et al. The elements of statistical learning. Vol. 2. No. 1. New York: springer, 2009.
4. Montgomery, Douglas C., and George C. Runger, Applied statistics and probability for engineers. John Wiley & Sons, 2010

BDS207 - Python Programming Lab

Total Marks : 100		**Max. Time : 2 Hrs.**
End Sem Exam: 60 Marks	**Internal Assessment: 40 Marks**	**Credits:- 2**
Min. Pass Marks : 50%	**Min. Pass Marks : 50%**	**Lectures:- 30**

Learning Objectives:

1. To impart the basic concepts of data structures and algorithms.
2. To understand concepts about searching and sorting techniques
3. To Understand basic concepts about stacks, queues, lists, trees and graphs
4. To understand the algorithms and develop the step by step solutions of problems with the help of data structures.

Learning Outcomes: On successful completion of this course, the students are able:

4. To analyze algorithms and their correctness.
5. To implement various searching and sorting techniques in different problems.
6. To have knowledge of concepts related with tree and graphs and able to apply them.

LIST OF EXERCISES:

1. Editing and executing Programs involving Flow Controls.
2. Editing and executing Programs involving Functions.
3. Program in String Manipulations
4. Creating and manipulating a Tuple
5. Creating and manipulating a List
6. Creating and manipulating a Dictionary
7. Object Creation and Usage
8. Program involving Inheritance
9. Program involving Overloading
10. Reading and Writing with Text Files and Binary Files
11. Combining and Merging Data Sets
12. Program involving Regular Expressions
13. Data Aggregation and GroupWise Operations

B.Sc. Data Science
Semester-III

BDS301 – Optimization Techniques

Total Marks : 100		**Max. Time : 3 Hrs.**
End Sem Exam: 60 Marks **Min. Pass Marks : 50%**	**Internal Assessment: 40 Marks** **Min. Pass Marks : 50%**	**Credits:- 3** **Lectures:- 45**

Learning Objectives:
The objective of this course is to teach various optimization techniques and highlight its usage in data sciencerelated applications.

Learning Outcomes: After completion of this course successfully the students will be able to:

1. Understand the optimization problems.
2. Solve the optimization problems.
3. Understand the use genetic algorithms for solving optimization problems.
4. Find the usage the optimization algorithms for data science tasks.

SECTION-A (15L)

Basics of optimization : Basics of optimization —how to formulate the problem, Maxima, minima, convex function,

global solution Linear programming, simplex algorithm, Integer programming, Constraint programming, Knapsackproblem,

SECTION-B (15L)

Randomized optimization : Randomized optimization techniques—hill climbing, simulated annealing,

SECTION-C (15L)

Introduction Genetic algorithms : Foundation of Evolutionary theory, Evolutionary Strategies, Evolutionary

programming,Evolutionary Algorithms, Evolutionary Algorithm Case Study, Genetic Algorithm, Genetic Representations, Initial Population, Fitness Function, Selection and Reproduction,

SECTION-D (15L)

Genetic Operators : Genetic Operators(Selection, Crossover, Mutation), Artificial Immune Systems, Other

Algorithms Harmony Search, Honey-Bee Optimization, Memetic Algorithms, Co-evolution, MultiObjectiveOptimization, Artificial Life, Constraint Handling

Recommended Books:

1. Optimization Techniques Hardcover,New Age Science Ltd; 1st edition (30 April 2009) by Chander Mohan, Kusum Deep.
2. Optimization Techniques: An Introduction,L. R. Foulds,Springer-Verlag.
3. Optimization Techniques,Chander Mohan and Kusum Deep,New Age Science.
4. Genetic Algorithms in Search, Optimization & Machine Learning, David E. Goldberg,Pearson Education India; 1st edition(1 December 2008)
5. Genetic Algorithms: Concepts and Designs (Advanced Textbooks in Control and Signal Processing), Kim-Fung Man, Kit-Sang Tang, Sam Kwong, Springer.

BDS302 - Database Management System (DBMS)

Total Marks : 100		**Max. Time : 3 Hrs.**
End Sem Exam: 60 Marks	**Internal Assessment: 40 Marks**	**Credits:- 3**
Min. Pass Marks : 50%	**Min. Pass Marks : 50%**	**Lectures:- 45**

Learning Objectives:

The objective of the course is to present an introduction to database management systems, with an emphasis on how toorganize, maintain and retrieve - efficiently, and effectively - information from a DBMS.

Learning Outcomes:

Upon successful completion of this course, students should be able to:

1. Describe the fundamental elements of database management systems.
2. Improve the database design by normalization.
3. Explain the basic concepts of relational data model, entity-relationship model, relational database design, relationalalgebra and SQL.
4. Design ER-models to represent simple database application scenarios
5. Convert the ER-model to relational tables, populate relational database and formulate SQL queries on data.
6. Familiar with basic database storage structures and access techniques: file and page organizations, indexing methodsincluding B tree, and hashing.

SECTION-A (6+8=14L)

1. Introduction :

Characteristics of database approach, data models, database system architecture and data independence.

2. Entity Relationship(ER) Modeling :

Entity types, relationships, constraints.

SECTION-B (20L)

3. Relation data model :

Relational model concepts, relational constraints, relational algebra, SQL queries

SECTION-C (15L)

4. Database design :

Mapping ER/EER model to relational database, functional dependencies, Lossless decomposition,Normal forms (up to BCNF).

SECTION-D (3+8=11L)

5. Transaction Processing :

ACID properties, concurrency control

6. File Structure and Indexing ;

Operations on files, File of Unordered and ordered records, overview of File organizations, Indexing structures for files(Primary index, secondary index, clustering index), Multilevel indexing using B and B+ trees.

Recommended Books:

1. R. Elmasri, S.B. Navathe, Fundamentals of Database Systems 6th Edition, Pearson Education, 2010.
2. R. Ramakrishanan, J. Gehrke, Database Management Systems 3rd Edition, McGraw-Hill, 2002.
3. A. Silberschatz, H.F. Korth, S. Sudarshan, Database System Concepts 6th Edition, McGraw Hill, 2010.
4. R. Elmasri, S.B. Navathe Database Systems Models, Languages, Design and application Courseming, 6th Edition, PearsonEducation,2013.

BDS303- Machine Learning

Total Marks : 100		**Max. Time : 3 Hrs.**
End Sem Exam: 60 Marks	**Internal Assessment: 40 Marks**	**Credits:- 3**
Min. Pass Marks : 50%	**Min. Pass Marks : 50%**	**Lectures:- 45**

Learning Objectives:
The objective of this course is to introduce machine learning fundamentals to students. This course provides introductory concepts of various machine learning techniques to students which will help to build foundation for further understanding. This course also aims to provides details of various steps involved in machine learning pipeline such as data collection, pre- processing, feature engineering etc. This course also introduce popular tools used in the area of machine learning.

Learning Outcomes: After completion of this course successfully the students will be able to:

1. Understand the various processes involve in machine learning.
2. Perform data cleaning and pre-processing
3. Decide and classify the problem as classification, prediction or clustering
4. Train and test machine learning algorithms

SECTION-A (5+5+5=15L)

Introduction to Machine learning :
Supervised and Unsupervised Learning.
Getting and Cleaning Data : Obtaining data from the web, from APIs, from databasesand from colleagues in various formats. basics of data cleaning and making data ―tidy.
Data pre-processing : Descriptive Data Summarization, Data Cleaning, Data Integration andTransformation, Data Reduction, Data Discretization and Concept Hierarchy Generation.

SECTION-B (15L)

Association Rule :
Mining Frequent Patterns, Associations, and Correlations: Basic Concepts and a Road Map, Association Rules, theApriori Algorithm Classification and Prediction

SECTION-C (15L)

Classification :
Classification, Issues Regarding Classification, Classification by Decision Tree Induction, Bayesian Classification,Rule-Based Classification, Metrics for Evaluating Classifier Performance, Holdout Method and Random Sub sampling

SECTION-D (5+10=15L)

Prediction : Prediction, Issues Regarding Prediction, Accuracy and Error Measures, Evaluating the Accuracy of a Classifier or Predictor.
Clustering : Cluster Analysis, Agglomerative versus Divisive Hierarchical Clustering,Distance Measures in Algorithmic, Evaluation of Clustering.
Tools and Frameworks : Scikit-learn, Weka and RStuido
Recommended Books:

1.Shalev-Shwartz, Shai, and Shai Ben-David. Understanding machine learning: From theory to algorithms. Cambridgeuniversity press, 2014.

2. Duda, Richard O., Peter E. Hart, and David G. Stork. Pattern classification. John Wiley & Sons, 2012.

3. Witten, Ian H., et al. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2016.

BDS305-Optimization Techniques(Python)-LAB

Total Marks : 100		**Max. Time : 2 Hrs.**
End Sem Exam: 60 Marks **Min. Pass Marks : 50%**	**Internal Assessment: 40 Marks** **Min. Pass Marks : 50%**	**Credits:- 2** **Lectures:- 30**

Learning Objectives:

The aim of this course is to introduce the fundamental concepts of Artificial Intelligence to students. The course will explain various important concepts such as searching techniques, Knowledge representation, Uncertainty and Natural Language Processing.

Learning Outcomes:

- Student will be able to understand different types of problem solving techniques.
- Student will be able to understand and use various searching techniques.
- Student will be aware to logic and knowledge representation techniques.

BDS306-DBMS – LAB

Total Marks : 100		Max. Time : 2 Hrs.
End Sem Exam: 60 Marks Min. Pass Marks : 50%	Internal Assessment: 40 Marks Min. Pass Marks : 50%	Credits:- 2 Lectures:- 30

Learning Objectives:

The aim of this course is to introduce the fundamental concepts of Artificial Intelligence to students. The course will explain various important concepts such as searching techniques, Knowledge representation, Uncertainty and Natural Language Processing.

Learning Outcomes:

- Student will be able to understand different types of problem solving techniques.
- Student will be able to understand and use various searching techniques.
- Student will be aware to logic and knowledge representation techniques.

Experiments list

1. Creating a database
2. Creating a table
3. Inserting records in a table
4. Altering the table structure.
5. Deleting data from table
6. Updating data from table.
7. Select command
8. Where clause
9. Aggregate functions
10. Numeric functions (Absolute, ceiling, floor, modulo, round off, square, Square Root, power)
11. Constraints
12. Group By, Having
13. Operators (and, or, not between, In , not in, is null, is not null, like, Order By)
14. String Functions (Lower, Upper, Replace, left-trim, right-trim, substring, Length, rename)
15. Drop (table, database)
16. Truncate
17. Sub Queries, Alias

BDS307-Machine Learning Lab

Total Marks : 100		Max. Time : 2 Hrs.
End Sem Exam: 60 Marks Min. Pass Marks : 50%	Internal Assessment: 40 Marks Min. Pass Marks : 50%	Credits:- 2 Lectures:- 30

Learning Objectives:

The aim of this course is to introduce the fundamental concepts of Artificial Intelligence to students. The course will explain various important concepts such as searching techniques, Knowledge representation, Uncertainty and Natural Language Processing.

Learning Outcomes:

- Student will be able to understand different types of problem solving techniques.
- Student will be able to understand and use various searching techniques.
- Student will be aware to logic and knowledge representation techniques.

List Of Experiments:

1. Write a program to implement the naïve Bayesian classifier for a sample training data set stored as a .CSV file. Compute the accuracy of the classifier, considering few test data sets.
2. Assuming a set of documents that need to be classified, use the naïve Bayesian algorithm.
3. Classifier model to perform this task. Built-in Java classes/API can be used to write the program. Calculate the accuracy, precision, and recall for your data set.
4. Write a program to implement k-Nearest Neighbour algorithm to classify the iris. print both correct and wrong predictions. Java/Python ML library classes can be used for this problem.
5. Write a program to implement Logistic Regression algorithm to classify the housing price data set. Print both correct and wrong predictions. Java/Python ML library classes can be used for this problem. 6. Write a program to implement and compare SVM, KNN and Logistic regression algorithm to classify the iPhone purchase records data set. Print both correct and wrong predictions. Java/ Python ML library classes can be used for this problem.

B.Sc. Data Science
Semester-IV

BDS401-Deep Learning

Total Marks : 100		**Max. Time : 3 Hrs.**
End Sem Exam: 60 Marks	**Internal Assessment: 40 Marks**	**Credits:- 3**
Min. Pass Marks : 50%	**Min. Pass Marks : 50%**	**Lectures:- 45**

Learning Objectives:

The objective of this course is to provide advance knowledge of machine learning techniques. This course mainly focused on Regression and Neural network based Machine learning algorithms. This aim to make students aware of various recent developments in the field of Neural network such as deep learning.

Learning Outcomes: After completion of this course successfully the students will be able to:

1. Perform regression analysis
2. Use to use Neural Network based model for classification and other task
3. Use to train and test deep learning based model for various tasks.
4. Use Python for building Deep learning based applications

SECTION-A (8+7=15L)

Linear Regression : Prediction using Linear Regression, Gradient Descent, Linear Regression with one variable, Linear Regression with multiple variables, Polynomial Regression, Feature Scaling/Selection.

Logistic Regression : Classification using Logistic Regression, Logistic Regression vs. Linear Regression, Logistic Regression with one variable and with multiple variables.

SECTION-B (15L)

Regularization : Regularization and its utility: The problem of Overfitting, Application of Regularization in Linear and Logistic Regression, Regularization and Bias/Variance.

SECTION-C (15L)

Neural Networks : Introduction, Model Representation, Gradient Descent vs. Perceptron Training, Stochastic Gradient Descent, Multilayer Perceptrons, Multiclass Representation, Backpropagation Algorithm.

SECTION-D (10+5=15L)

Deep Learning :

History, Scope and specification, why deep learning now, building block of neural network, neural networks, Deep learning hardware. Feedforward neural networks, xor model, cost function estimation (maximum likelihood), units, activation functions, layers, , normalization, hyper-parameter tuning, Convolution neural networks, architecture, recurrent neural networks, architecture, types and overview, GAN (Generative Adversarial Networks).

Deep learning applications :

Computer vision, sentiment analysis, music generation, text generation, neural style transfer, image captioning

Recommended Books:

1. Ethem Alpaydin, "Introduction to Machine Learning" 2nd Edition, The MIT Press, 2009.
2. Tom M. Mitchell, "Machine Learning", First Edition by Tata McGraw-Hill Education, 2013.
3. Christopher M. Bishop, "Pattern Recognition and Machine Learning" by Springer, 2007.
4. Mevin P. Murphy, "Machine Learning: A Probabilistic Perspective" by The MIT Press, 2012.

BDS402-IoT Programming and Big Data

Total Marks : 100		**Max. Time : 3 Hrs.**
End Sem Exam: 60 Marks	**Internal Assessment: 40 Marks**	**Credits:- 3**
Min. Pass Marks : 50%	**Min. Pass Marks : 50%**	**Lectures:- 45**

Learning Objectives:
The objective of this course is to know and appreciate the software needs of an IoT project and understand how data is managed in an IoT network. This course course also aim to explain how to apply software solutions for different systems and Big Data to IoT concept designs. This course focus on Python to write scripts to manage large data files collected from sensor data and interact with the real world via actuators and other output devices.

Learning Outcomes: After completion of this course successfully the students will be able to:

1. Find the applications of IoT in real world and use techniques to build software.
1. Understand the IoT network and sensor data
2. Collect and analyze large data collect through various sensor
3. Use Python for building IoT and big data based applications.

SECTION-A (15L)

Introduction to Big Data from the IoT :
Develop an understanding of the data generated by IoT, and how it is collected; Recognise the problems involved with gathering data and some approaches for addressing these problems; Gain an overview of data storage

SECTION-B (15L)

Data at the Edge :
Understand the process of data acquisition; Be able to analyse where to process data using Edge, Fog or Cloud; Understand how, when, and where to bundle and store IoT data

SECTION-C (15L)

Data in the Cloud :

Understand the storage, analysis and cleaning of data; Understand why data is stored and processed in the Cloud; Appreciate the costs and benefits of live data versus stored data.

SECTION-D (15L)

Obtaining, Visualising and Analysing Data :
Understand some methods for cleaning, summarising and visualising a large dataset; Construct and use a simple predictive model for predicting the location of a device using signal strength and orientation.

Learn how to use Python, R and RStudio to performance analysis of a large dataset;
Case studies and projects

Recommended Books:

1. Internet of Things – A Hands-on Approach, Arshdeep Bahga and Vijay Madisetti, Universities Press, 2015, ISBN: 9788173719547
2. Getting Started with Raspberry Pi, Matt Richardson & Shawn Wallace, O'Reilly (SPD), 2014, ISBN: 9789350239759

BDS403 - Data warehouse and Data Mining

Total Marks : 100		**Max. Time : 3 Hrs.**
End Sem Exam: 60 Marks	**Internal Assessment: 40 Marks**	**Credits:- 3**
Min. Pass Marks : 50%	**Min. Pass Marks : 50%**	**Lectures:- 45**

Learning Objectives:

1. To understand data warehouse concepts, architecture, business analysis and tools
2. To understand data pre-processing and data visualization techniques
3. To study algorithms for finding hidden and interesting patterns in data
4. To understand and apply various classification and clustering techniques using tools.

Learning Outcomes: After completion of this course successfully the students will be able to:

1. Design a Data warehouse system and perform business analysis with OLAP tools.
2. Apply suitable pre-processing and visualization techniques for data analysis
3. Apply frequent pattern and association rule mining techniques for data analysis
4. Apply appropriate classification and clustering techniques for data analysis

SECTION-A (5+5+5=15L)

Data Warehouse Fundamentals : Introduction to Data Warehouse, OLTP Systems; Differences between OLTP Systems and Data Warehouse: Characteristics of Data Warehouse; Functionality of Data Warehouse: Advantages and Applications of Data Warehouse; Advantages, Applications: Top - Down and Bottom-Up Development Methodology: Tools for Data warehouse development: Data Warehouse Types:

Planning and Requirements : Introduction: Planning Data Warehouse and Key Issues: Planning and Project Management in constructing Datawarehouse: Data Warehouse Project; Data Warehouse Development Life Cycle, Kimball Lifecycle Diagram, Requirements Gathering Approaches: Team organization, Roles, and Responsibilities:

Data Warehouse Architecture : Introductions, Components of Data Warehouse Architecture: Technical Architectures; Data warehouse architectures 1: Data warehouse architecture 2: Data warehouse architecture 3: Tool selection: Federated Data Warehouse Architecture:

SECTION-B (5+5+5=15L)

Dimensional Modeling : Introduction: E-R Modeling: Dimensional Modeling: E-R Modeling VS Dimensional Modeling: Data Warehouse Schemas; Star Schema, Inside Dimensional Table, Inside Fact Table, Fact Less Fact Table, Granularity, Star Schema Keys: Snowflake Schema: Fact Constellation Schema:

Extract, Transform and Load : Introduction: ETL Overview or Introduction to ETL: ETL requirements and steps: Data Extraction; Extraction Methods, Logical Extraction Methods, Physical Extraction Methods: Data Transformation; Basic Tasks in Transformation, Major Data Transformation Types: Data loading; Data Loading Techniques: ETL Tools:

Data Warehouse & OLAP : Introduction: What is OLAP?; Characteristics of OLAP, Steps in the OLAP Creation Process, Advantageous of OLAP: What is Multidimensional Data: OLAP Architectures; MOLAP,

SECTION-C (5+5+5=15L)

Meta data Management in Data Warehouse : Introductions to Metadata: Categorizing Meta data: Metadata management in practice; Meta data requirements gathering, Meta data classification, Meta data collection strategies: Meta Data Management in Oracle and SAS: Tools for Meta data management:

Introduction to Data Mining : Introduction: Scope of Data Mining: What is Data Mining; How does DataMining Works, Predictive Modeling: Data Mining and Data Warehousing: Architecture for D ata Mining: Profitable Applications: Data Mining Tools:

Business Intelligence : Introduction, Business Intelligence, Business Intelligence tools, Business Intelligence Infrastructure, Business Intelligence Applications, BI versus Data Warehouse, BI versus Data Mining, Future of BI.

SECTION-D (3+4+4+4=15L)

Data Pre-processing : Introduction, Data Preprocessing Overview, Data Cleaning, Data Integration and Transformation, Data Reduction, Discretization and Concept Hierarchy Gene ration.

Data Mining Techniques- An Overview : Introduction, Data Mining, Data Mining Versus Database Management System, Data Mining Techniques- Association rules, Classification, Regression, Clustering, Neural networks.

Clustering : Introduction, Clustering, Cluster Analysis, Clustering Methods- K means, Hierarchical clustering, Agglomerative clustering, Divisive clustering, clustering and segmentation software, evaluating clusters.

Web Mining : Introduction, Terminologies, Categories of Web Mining – Web Content Mining, WebStructure Mining, Web Usage Mining, Applications of Web Mining, and Agent based and Data base approaches, Web mining Software. Applications of Data mining.

Recommended Books:

1. Alex Berson and Stephen J.Smith, ―Data Warehousing, Data Mining & OLAP‖, Tata McGraw – HillEdition, 35th Reprint 2016.
2. K.P. Soman, Shyam Diwakar and V. Ajay, ―Insight into Data Mining Theory and Practice, EasternEconomy Edition, Prentice Hall of India, 2006.
3. Ian H.Witten and Eibe Frank, ―Data Mining: Practical Machine Learning Tools and Techniques, Elsevier,Second Edition.

BDS404- Deep Learning-Lab

LIST OF EXERCISES:

1. Setting up the Spyder IDE Environment and Executing a Python Program
2. Installing Keras, Tensorflow and Pytorch libraries and making use of them
3. Artificial Neural Networks
4. Convolutional Neural Networks
5. Image Transformations
6. Image Gradients and Edge Detection
7. Image Contours
8. Image Segmentation
9. Harris Corner Detection
10. Face Detection using Haar Cascades
11. Chatbot Creation

BDS405-IoT Programming and Big Data LAB

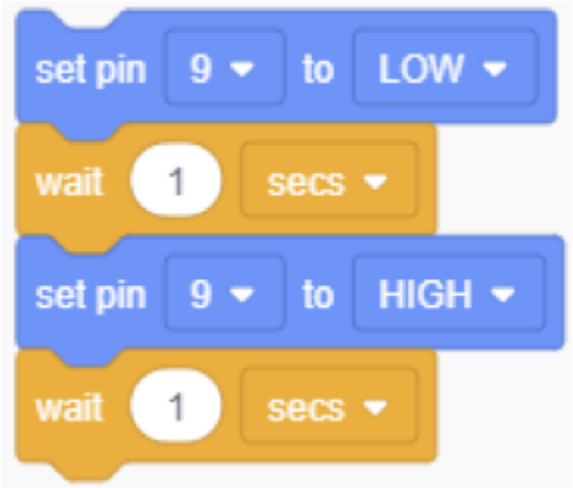
Practical task 1 requires that you program an Arduino in Tinkercad® Circuits to respond to a switch, read from a sensor and write to a multi-coloured LED.

You should already have created your own Tinkercad® account and become familiar with creating and using circuit simulations during the activities in this course. If not, go back and do them before beginning this assessment.

Please **download and follow** the instructions below and then return here to complete the questions.

Download instructions - IoT4x_Module1_PracticalTask1.pdf

Question 2 - Screenshots

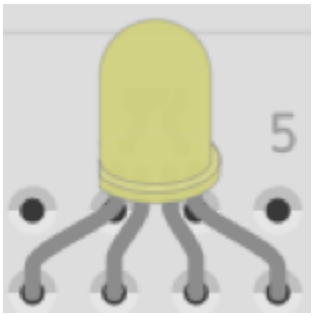
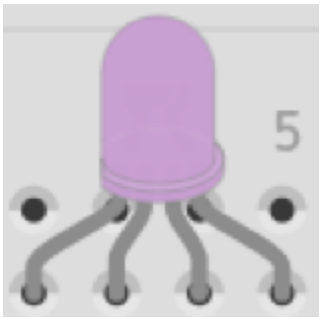
Screenshot 1	**Screenshot 2**
	
Screenshot 3	**Screenshot 4**

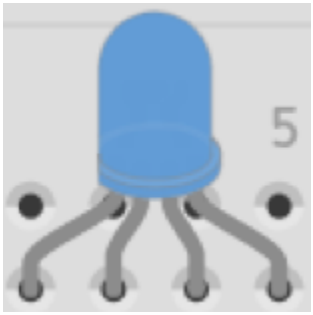
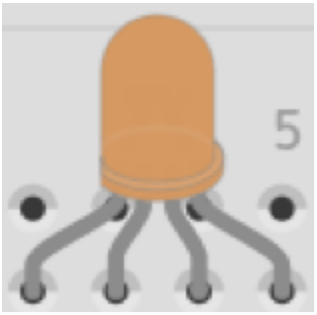
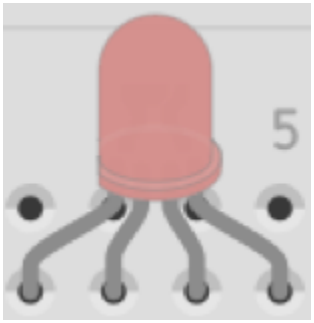
Question 3 - Information

Once you have successfully programmed the Arduino to the requirements specified, start the simulation and set the temperature on the sensor to:

- -40°C,
- then 50°C,
- then 125°C

and take notice of the **colour** of the LED each time.

Yellow	Light Blue	Pink
		

Dark Blue	Orange	Red
		

By the end of this module you will:

- Understand the process of data acquisition;
- Be able to analyse where to process data using Edge, Fog or Cloud;
- Understand how, when and where to bundle and store IoT data.

Case studies:

1. Cow tracking and monitoring on a dairy farm
2. Traffic management
3. Space utilisation on a university campus

BDS406- Data warehouse and Data Mining – LAB

Learning Objectives:

1. Learn how to build a data warehouse and query it (using open source tools like Pentaho Data Integration Tool, Pentaho Business Analytics).

2. Learn to perform data mining tasks using a data mining toolkit (such as open source WEKA).

3. Understand the data sets and data preprocessing.

4. Demonstrate the working of algorithms for data mining tasks such association rule mining, classification, clustering and regression.

5. Exercise the data mining techniques with varied input values for different parameters.

6. To obtain Practical Experience Working with all real data sets.

7. Emphasize hands-on experience working with all real data sets.

Learning Outcomes:

1. Ability to understand the various kinds of tools.

2. Demonstrate the classification, clustering and etc. in large data sets.

3. Ability to add mining algorithms as a component to the exiting tools.

4. Ability to apply mining techniques for realistic data

1. Unit-I Build Data Warehouse and Explore WEKA

2 Unit-II Perform data preprocessing tasks and Demonstrate performing association rule mining on data sets

3 Unit-III Demonstrate performing classification on data sets

4 Unit-IV Demonstrate performing clustering on data sets

5 Unit-V Demonstrate performing Regression on data sets

6 Task 1: Credit Risk Assessment. Sample Programs using German Credit Data

7 Task 2: Sample Programs using Hospital Management System

8 Beyond the Syllabus -Simple Project on Data Preprocessing

A. Build Data Warehouse/Data Mart (using open source tools like Pentaho Data Integration Tool, Pentaho Business Analytics; or other data warehouse tools like Microsoft-SSIS,Informatica,Business Objects,etc.,)

A.(i) Identify source tables and populate sample data. The data warehouse contains 4 tables:

1. Date dimension: contains every single date from 2006 to 2016.

2. Customer dimension: contains 100 customers. To be simple we'll make it type 1 so we don't create a new row for each change.

3. Van dimension: contains 20 vans. To be simple we'll make it type 1 so we don't create a new row for each change.

4. Hire fact table: contains 1000 hire transactions since 1st Jan 2011. It is a daily snapshot fact table so that every day we insert 1000 rows into this fact table. So over time we can track the changes of total bill, van charges, satnav income, etc.

Create the source tables and populate them So now we are going to create the 3 tables in HireBase database: Customer, Van, and Hire. Then we populate them.

B.Sc. Data Science
Semester-V
BDS501- Big Data Analytics through Spark

Total Marks : 100		**Max. Time : 3 Hrs.**
End Sem Exam: 60 Marks	**Internal Assessment: 40 Marks**	**Credits:- 3**
Min. Pass Marks : 50%	**Min. Pass Marks : 50%**	**Lectures:- 45**

Learning Objectives:
The objective of this course is to know and appreciate the software needs of an IoT project and understand how datais managed in an IoT network. This course course also aim to explain how to apply software solutions for different systems and Big Data to IoT concept designs. This course focus on Python to write scripts to manage large data files collected from sensor data and interact with the real world via actuators and other output devices.

Learning Outcomes: After completion of this course successfully the students will be able to:

4. Find the applications of IoT in real world and use techniques to build software.
5. Understand the IoT network and sensor data
6. Collect and analyze large data collect through various sensor
7. Use Python for building IoT and big data based applications.

Unit – I: Introduction to Spark Apache Spark Ecosystem - Setting up the Spark Python Environment – Execution of a PySpark Program – Resilient Distributed Datasets – Spark Architecture – Spark Project Workflow.

Unit – II: Spark Programming with Python Loading and Storing Data – Transformations – Actions – Key-Value Resilient Distributed Datasets – Local Variables – Broadcast Variables – Accumulators – Partitioning – Persistence.

Unit – III: Spark SQL Overview of Spark SQL – Spark Session – Data Frames – Schema of a Data Frame – Operations supported by Data Frames – Filter, Join, GroupBy, Agg operations – Nesting the Operations – Temporary Tables – Viewing and Querying Temporary Tables.

Unit – IV: Spark Streaming Use Cases for Realtime Analytics – Transferring, Summarizing, Analysing Realtime data – Data Sources supported by Spark Streaming – Flat files, TCP/IP – Flume – Kafka – Kinesis – Streaming Context – DStreams – Dstream RDDs – Dstream Processing.

Unit – V: Machine Learning with Spark Linear Regression – Decision Tree Classification – Principal Component Analysis – Random Forest Classification – Text Pre-processing with TF-IDF – Naïve Bayes Classification – KMeans Clustering – Recommendation Engines.

REFERENCES:

1. Tomasz Drabos, “Learning PySpark”, PACKT, 2017.
2. Padma Priya Chitturi, “Apache Spark for Data Science”, PACKT, 2017.
3. Holden Karau, “ Learning Spark”. PACKT, 2016.
4. Sandy Riza, “Advanced Analytics with Spark”, O’ Reilly, 2016.
5. Romeo Kienzler, “Mastering Apache Spark”, PACKT, 2017.

BDS502-Introduction to Artificial Intelligence

Total Marks : 100		**Max. Time : 3 Hrs.**
End Sem Exam: 60 Marks	**Internal Assessment: 40 Marks**	**Credits:- 3**
Min. Pass Marks : 50%	**Min. Pass Marks : 50%**	**Lectures:- 45**

Learning Objectives:

The aim of this course is to introduce the fundamental concepts of Artificial Intelligence to students. The course will explain various important concepts such as searching techniques, Knowledge representation, Uncertainty andNatural Language Processing.

Learning Outcomes:

- Student will be able to understand different types of problem solving techniques.
- Student will be able to understand and use various searching techniques.
- Student will be aware to logic and knowledge representation techniques.

SECTION-A (6+5=11L)

1. Introduction:

Introduction to Artificial Intelligence, Background and Applications, Turing Test and RationalAgent approaches to AI, Introduction to Intelligent Agents, their structure, behavior and environment.

Problem Solving: Problem Characteristics, Production Systems, Control Strategies

SECTION-B (15L)

Searching Techniques : Breadth First Search, Depth First Search, Hill climbing and its Variations, Heuristics Search Techniques: Best First Search, A* algorithm, Constraint Satisfaction Problem, Means-End Analysis, Introduction to Game Playing, Min-Max and Alpha-Beta pruning algorithms.

SECTION-C (20L)

3. **Knowledge Representation :**Introduction to First Order Predicate Logic, Resolution Principle, Unification, Semantic Nets, Conceptual Dependencies, Frames, and Scripts, Production Rules, Conceptual Graphs. Courseming in Logic (PROLOG)

SECTION-D (8+6=14L)

4. **Dealing with Uncertainty and Inconsistencies :** Truth Maintenance System, Default Reasoning, Probabilistic Reasoning, Bayesian Probabilistic Inference, Possible World Representations.

5. **Understanding Natural Languages :**Parsing Techniques, Context-Free and Transformational Grammars, Recursive andAugmented Transition Nets.

Recommended Books:

1. DAN.W. Patterson, Introduction to A.I and Expert Systems – PHI, 2007.
2. Russell &Norvig, Artificial Intelligence-A Modern Approach, LPE, Pearson Prentice Hall,2nd edition, 2005.
3. Rich & Knight, Artificial Intelligence – Tata McGraw Hill, 2nd edition, 1991.
4. W.F. Clocksin and Mellish, Courseming in PROLOG, Narosa Publishing, House, 3rd edition,
5. Ivan Bratko, Prolog Courseming for Artificial Intelligence, Addison-Wesley, Pearson 3rd ed

BDS503-Machine Learning Operations (ML Ops)

Total Marks : 100		Max. Time : 3 Hrs.
End Sem Exam: 60 Marks Min. Pass Marks : 50%	Internal Assessment: 40 Marks Min. Pass Marks : 50%	Credits:- 3 Lectures:- 45

Learning Objectives:

The aim of this course is to introduce the fundamental concepts of Artificial Intelligence to students. The course will explain various important concepts such as searching techniques, Knowledge representation, Uncertainty andNatural Language Processing.

Learning Outcomes:

- Student will be able to understand different types of problem solving techniques.
- Student will be able to understand and use various searching techniques.
- Student will be aware to logic and knowledge representation techniques.

Unit 1:MLOps Fundamentals, Why and When do we need MLOps, Data Scientists' Pain Points, Machine Learning Lifecycle, MLOps Architecture and TensorFlow Extended Components, Why and When to Employ MLOps.

Unit 2:Understanding the Main Kubernetes Components:

Introduction to Containers, Containers and Container Images, Introduction to Kubernetes, Introduction to Google Kubernetes Engine, Compute Options Detail, Kubernetes Concepts, The Kubernetes Control Plane, Google Kubernetes Engine Concepts, Deployments, Ways to Create Deployments, Services and Scaling, Updating Deployments, Rolling Updates, Blue-Green Deployments, Canary Deployments, Managing Deployments, Jobs and CronJobs, Parallel Jobs.

Unit 3: Introduction to Containers, Containers and Container Images, Introduction to Kubernetes, Introduction to Google Kubernetes Engine, Containers and Kubernetes in Google Cloud, Kubernetes Concepts, The Kubernetes Control Plane, Google Kubernetes Engine Concepts, Deployments, Updating Deployments, Jobs.

Unit 4: Introduction to AI Platform Pipelines: Overview, Introduction to AI Platform Pipelines, Concepts, When to use, Ecosystem, Getting Started with Google Cloud and Qwiklabs.

Unit 5:Training, Tuning and Serving on AI Platform:

System and concepts overview, Create a reproducible dataset, Implement a tunable model, Build and push a training container, Train and tune the model, Serve and query the model.

Unit 6:Kubeflow Pipelines on AI Platform: System and concept overview, Describing a Kubeflow Pipeline with KF DSL, Pre-built components, Lightweight Python Components, Custom components, Compile, upload and Run.

CI/CD for Kubeflow Pipelines on AI Platform: Concept Overview, Cloud Build Builders, Cloud Build Configuration, Cloud Build Triggers.

BDS504: Elective – I

BDS505: BIG DATA ANALYTICS THROUGH SPARK - LAB

Total Marks : 100		**Max. Time : 2 Hrs.**
End Sem Exam: 60 Marks **Min. Pass Marks : 50%**	**Internal Assessment: 40 Marks** **Min. Pass Marks : 50%**	**Credits:- 2** **Lectures:- 30**

LIST OF EXERCISES:

1. Program involving Resilient Distributed Datasets
2. Program involving Transformations and Actions
3. Program involving Key-Value Resilient Distributed Datasets
4. Program involving Local Variables, Broadcast Variables and Accumulators
5. Program involving Filter, Join, GroupBy, Agg operations
6. Viewing and Querying Temporary Tables
7. Transferring, Summarizing and Analysing Twitter data
8. Program involving Flume, Kafka and Kinesis
9. Program involving DStreams and Dstream RDDs
10. Linear Regression
11. Decision Tree Classification
12. Principal Component Analysis
13. Random Forest Classification
14. Text Pre-processing with TF-IDF
15. Naïve Bayes Classification
16. K-Means Clustering

BDS506: Artificial Intelligence (PROLOG / Python) -LAB

Total Marks : 100		**Max. Time : 2 Hrs.**
End Sem Exam: 60 Marks **Min. Pass Marks : 50%**	**Internal Assessment: 40 Marks** **Min. Pass Marks : 50%**	**Credits:- 2** **Lectures:- 30**

Learning Objectives/Outcomes:

After successful completion of this course, the student should be able to:

1. Understand and use problem searching agents (informed and uninformed methods)
2. Understand and use game playing techniques
3. Understand and use agents that reason logically
4. Understand and use building knowledge bases and theorem proving
5. Understand and use uncertainty and probabilistic reasoning

List of Experiments:

1. Churn Analysis and Prediction (Survival Modelling)

- Cox-proportional models
- Churn Prediction

2. Credit card Fraud Analysis

- Imbalanced Data
- Neural Network

3. Sentiment Analysis or Topic Mining from New York Times

- Part-of-Speech Tagging
- Stemming and Chunking

4. Sales Funnel Analysis

- A/B testing
- Campaign effectiveness, Web page layout effectiveness
- Scoring and Ranking

5. Recommendation Systems and Collaborative filtering

- User based
- Item Based
- Singular value decomposition–based recommenders

6. Customer Segmentation and Value

- Segmentation Strategies
- Lifetime Value

7. Portfolio Risk Conformance

- Risk Profiling
- Portfolio Optimization

8. Uber Alternative Routing

- Graph Construction
- Route Optimization

BDS507: Project Work –Minor

Capstone Projects (Option to Bring Your Own Project)

1. Real-time system for Tweet Analytics
2. Food Image Segmentation
3. Talent Retention and Attrition Prediction
4. Identification of Quora question pairs with the same intent
5. Stock Market predictions based on Time Series
6. Prediction of Client Subscription to a Bank term Deposit
7. Direct Retail Marketing efforts based on Customer Segmentation using ML based Clustering techniques
8. Movie Recommendation System
9. Predict the future daily-demand for a large Logistics Company
10. Achieving image super-resolution using a Generative Adversarial Network
11. Determine key factors driving literacy rate in the Indian demography using Predictive Data Analytics
12. Urban Crime Data Analytics for safety improvement
13. Breast Cancer classification from digitized FNA image feature measurements
14. Exploratory and Predictive Data Analytics using Indian Premier League (IPL) dataset
15. Anomaly detection in Bearing Vibration Measurements

B.Sc. Data Science
Semester-VI

BDS601: NOSQL DATABASES

Total Marks : 100		**Max. Time : 3 Hrs.**
End Sem Exam: 60 Marks	**Internal Assessment: 40 Marks**	**Credits:- 3**
Min. Pass Marks : 50%	**Min. Pass Marks : 50%**	**Lectures:- 45**

- An Overview of NoSQL (1 hour)
- HDFS (3 hours)
- Apache Hive as an HDFS Data Warehouse (5 hours)
- HBase (5 hours)
- MongoDB (6 hours)
- Cassandra (7 hours)
- Neo4j (3 hours)

Unit – I: NoSQL and HDFS

Unit – II: Hive

Unit – III: HBase

Unit – IV: MongoDB Introduction – Features - Data types - Mongo DB Query language - CRUD operations – Arrays - Functions: Count – Sort – Limit – Skip – Aggregate - Map Reduce. Cursors – Indexes - Mongo Import – Mongo Export.

Unit – V: Cassendra Introduction – Features - Data types – CQLSH - Key spaces - CRUD operations – Collections – Counter – TTL - Alter commands - Import and Export - Querying System tables.

BDS602: CLOUD COMPUTING

Total Marks : 100		**Max. Time : 3 Hrs.**
End Sem Exam: 60 Marks	**Internal Assessment: 40 Marks**	**Credits:- 3**
Min. Pass Marks : 50%	**Min. Pass Marks : 50%**	**Lectures:- 45**

Unit – I:
Introduction Evolution of Cloud Computing –Essential Characteristics of cloud computing – Operational models such as private, dedicated, virtual private, community, hybrid and public cloud – Service models such as IaaS, PaaS and SaaS – Governance and Change Management – Business drivers, metrics and typical use cases. Example cloud vendors – Google cloud platform, Amazon AWS, Microsoft Azure, Pivotal cloud foundry and Open Stack.

Unit – II:
Infrastructure Services Basics of Virtual Machines - Taxonomy of Virtual Machines. Virtualization Architectures. Challenges with Dynamic Infrastructure - Principles of Infrastructure as Code - Considerations for Infrastructure Services and Tools - Monitoring: Alerting, Metrics, and Logging - Service Discovery - Server Provisioning via Templates - Patterns and Practices for Continuous Deployment - Organizing Infrastructure and Testing Infrastructure - Change Management Pipelines for Infrastructure.

Unit – III:
Platform Engineering Cloud Native Design and Microservices– Containerized - Dynamically orchestrated design – Continuous delivery - Support for a variety of client devices – Monolithic vs Microservices Architecture - Characteristics of microservice architecture – 12 factor application design - Considering service granularity – Scalable Services - Sharing dependencies between microservices - Stateless versus Stateful microservices - Service discovery – Service Registry – Performance Considerations.

Unit – IV:
Serverless Architecture and DevOps Function as a Service (FaaS) - Backend as a Service (BaaS) - Advantages of serverless architectures - Taking a hybrid approach to serverless architecture - Function deployment and Function invocation. Introduction to DevOps - The Deployment Pipeline - The Overall Architecture - Building and Testing - Deployment - Crosscutting Concerns such as Monitoring, Scalability, Repeatability, Reliability, Recoverability, Interoperability, Testability, and Modifiability.

Unit – V:
Cloud Security Considerations – STRIDE Threat Model - Cloud Security Challenges – Cloud specific Cryptographic Techniques – CIA Triad – Security by Design – Common Security Risks - Risk Management – Security Monitoring – Security Architecture Design – Data Security – Application Security – Virtual Machine Security.

REFERENCES:
1. Dr.AnandNayyar, (2019), “Handbook of Cloud Computing”, BPB
2. Mastering Azure Machine Learning, By Christoph Korner and Kaijisse Waaijer, April 2020
3. Hands-On Machine Learning on Google Cloud Platform,By Giuseppe Ciaburro, V Kishore Ayyadevara and Alexis Perrier, April 2018
4. Learning Path: AWS Certified Machine Learning-Specialty ML, By Noah Gift, April 2019
5. Software Architect's Handbook, by Joseph Ingeno, Published by Packt Publishing, 2018
6. Architecting Cloud Computing Solutions by Scott Goessling, Kevin L. Jackson, Publisher: Packt Publishing, Release Date: May 2018
7. Microservices: Flexible Software Architecture, by Eberhard Wolff, Publisher: Addison-Wesley Professional, Release Date: October 2016

BDS603: Big Data Acquisition and Analysis

Learning Objectives:

Learn to develop Hadoop applications for storing processing and analyzing data stored in Hadoop cluster.

The course is mainly covering Big Data tools for Data Transformation (Apache PIG), Data Analysis (HIVE) and for handling unstructured data HBase.

To Understand the complexity and volume of Big Data and their challenges.

To analyse the various methods of data collection.

To comprehend the necessity for pre-processing Big Data and their issues.

Learning Outcome:

1. Identify the various sources of Big Data
2. Able to collect and store Big Data from various sources
3. Able to write Pig Scripts- Extract, Transform and Load the data on HDFS
4. Able to write Hive Scripts- Extract, Transform, Load and Analyse the data present in HDFS
5. Able to write scripts to extract data from structured and un-structured data for analytics
6. Able to extract and process semi and un-structured data using HBase

Unit- I Introduction To Big Data Acquisition:

Big data framework – fundamental concepts of Big Data Management and analytics – Current challenges and trends in Big Data Acquisition. Map Reduce Algorithm- Hadoop Storage [HDFS], Common Hadoop Shell commands - Anatomy of File Write and Read, NameNode, Secondary NameNode, and DataNode - Hadoop Configuration – Pig Configuration – Hive Configuration - HBase Configuration.

Unit-II Data Collection And Transmission:

Big data collection – Strategies – Types of Data Sources – Structured Vs Unstructured data – ELT vs ETL – storage infrastructure requirements – Collection methods – Log files – sensors – Methods for acquiring network data (Libcap-based and zero-copy packet capture technology) – Specialized network monitoring softwares (Wireshark, Smartsniff and Winnetcap) – Mobile equipments, Transmission methods, Issues.

Unit-III Apache Pig –

Introduction - Pig features - Pig Architecture - Pig Execution modes, Pig Grunt shell and Shell commands. Pig Latin Basics: Data model, Data Types, Operators - Pig Latin Commands - Load & Store , Diagnostic Operators, Grouping, Cogroup, Joining, Filtering, Sorting, Splitting - Built-In Functions, User define functions. Pig Execution Modes: Batch Mode – Embedded Mode – Pig Execution in Batch Mode –Use cases - Map Reduce programs with Pig – Pig Vs SQL

Unit-IV Hive: Introduction - Hive Features - Hive architecture -Hive Meta store - Hive data types -51 Hive Tables - Table types - Creating database, Altering database, Create table, alter table, Drop table, Built-In Functions - Built-In Operators, User defined functions(UDFs), View, Pig Vs Hive. HiveQL– Introduction, HiveQL Select, HiveQL – MapReduce using HiveQL OrderBy, Group By Joins, LIMIT, Distribute By , Cluster By - Sorting And Aggregation – Partitioning: Static & Dynamic partitioning – Index Creation - Bucketing – Analysis of MapReduce execution – Hive Optimization – Setting Hiivng Parameters. Comparison between MapReduce, Hive QL and SQL. UseCase: Implementation of MapReduce programs with HiveQL.

Unit-V Hbase: HBasics, Features of HBase, Concepts, Clients, Example, Hbase Versus RDBMS, Limitations of HBase Big Data Privacy And Applications: Data Masking – Privately identified Information (PII) – Privacy preservation in Big Data – Popular Big Data Techniques and tools – ApplicationsSocial Media Analytics – Fraud Detection.

References

1. Bart Baesens, “Analytics in a Big Data World: The Essential Guide to Data Science and its Applications’, John Wiley & Sons, 2014.
2. Tom White “ Hadoop: The Definitive Guide” Third Edit on, O’reily Media, 2012.
3. Seema Acharya, Subhasini Chellappan, "Big Data Analytics" Wiley 2015.
4. Min Chen. Shiwen Mao, Yin Zhang. Victor CM Leung, Big Data: Related Technologies, Challenges and Future Prospects, Springer, 2014.
5. Michael Minelli, Michele Chambers Ambiga Dhiraj, “Big Data, Big Analytics : Emerging Business Intelligence and Analytic Trends”, John Wiley & Sons, 2013.
6. Raj. Pethuru “ Handbook of Research on Cloud Infrastructures for Big Data Analytics”, IGI Global.

Student Activity:

Case study I: “BankAmeriDeals” provides cash-back offers to credit and debit-card customers based upon analyses of their prior purchases.
Case Study II: GOOGLE: Working with the U.S. Centers for Disease Control, tracks when users are inputting search terms related to flu topics, to help predict which regions may experience outbreaks.
Case Study III: Twitter data Analysis RECOMMENDED CO-CURRICULAR ACTIVITIES: (Co-curricular activities shall not promote copying from textbook or from others work and shall encourage self/independent and group learning)

A. Measurable

1. Assignments (in writing and doing forms on the aspects of syllabus content and outside the syllabus content. Shall be individual and challenging)
2. Student seminars (on topics of the syllabus and related aspects (individual activity))
3. Quiz (on topics where the content can be compiled by smaller aspects and data (Individuals or groups as teams))
4. Study projects (by very small groups of students on selected local real-time problems pertaining to syllabus or related areas. The individual participation and contribution of students shall be ensured (team activity)

B. General

1. Group Discussion
2. Try to solve MCQ’s available online.
3. Others

RECOMMENDED CONTINUOUS ASSESSMENT METHODS:

Some of the following suggested assessment methodologies could be adopted;
1. The oral and written examinations (Scheduled and surprise tests)
2. Closed-book and open-book tests
3. Problem-solving exercises
4. Practical assignments and laboratory reports
5. Observation of practical skills
6. Individual and group project reports like “Movie Lens Data Analysis”, "Youtube Click stream Data Analysis, Twitter Data Analysis, etc
7. Efficient delivery using seminar presentations,
8. Viva voce interviews.
9. Computerized adaptive testing, literature surveys and evaluations,
10. Peers and self-assessment, outputs form individual and collaborative work

BDS604: Elective – II

BDS605: Big Data Acquisition and Analysis Lab

1. Hadoop Cluster Setup
   - Perform setting up and Installing Hadoop in its three operating modes:
     - o standalone
     - o Pseudo distributed
     - o fully distributed
   - Use web based tools to monitor your Hadoop setup.

2. Install and Run Pig and also use Pig Shell commands to display the list of files in HDFS

3. Install and Run Hive and also use Hive Shell commands to display the list of files in HDFS

4. Install and Run HBase and also use HBase Shell commands to display the version and user of HBase

5. Use Hive to create, alter, and drop databases, tables, views, functions, and indexes

6. Write and execute Pig Script to Load data into a Pig relation without a schema

7. Write and execute Pig Script Load data into a Pig relation with a schema

8. Write a Pig script to find the word count in a text file

9. Write a Pig Script that mines weather data (NCDC). Weather sensors collecting data every hour at many locations across the globe gather a large volume of log data, which is a good candidate for analysis with MapReduce, since it is semi structured and record-oriented.

Data available at: ftp://ftp.ncdc.noaa.gov/pub/data/noaa/.

- Find average, max and min temperature for each year in NCDC data set
- Filter the readings of a set based on value of the measurement, Output the line of input files associated with a temperature value greater than 30.0 and store it in a separate file.

10. Write HiveQL command to create Weather table and to find the year-wise maximum temperature

11. Write a Pig Script to remove null and duplicate values from the given input file.

12. Write Pig scripts to implement filter, project, sort, group by, joins

13. Write Hive Query to create database, managed table, external table, join, index, view, etc

14. Create a table in HBase and insert the data into with Shell

15. Display the data present in a HBase table using Shell

BDS606: NOSQL DATABASES - LAB

1. Exercises on HDFS
2. Exercises on Apache Hive as an HDFS Data Warehouse
3. Exercises on HBase
4. Exercises on MongoDB
5. Exercises on Cassandra
6. Exercises on Neo4j

BDS607: Project Work / Dissertation

Annexure –I: Syllabus for Elective-I papers

1. Technologies for Data Science

Total Marks : 100		**Max. Time : 3 Hrs.**
End Sem Exam: 60 Marks **Min. Pass Marks : 50%**	**Internal Assessment: 40 Marks** **Min. Pass Marks : 50%**	**Credits:- 3** **Lectures:- 45**

Learning Objectives:
This course is totally practical in nature and train students on various tools, frameworks, and libraries used for Big dataanalysis. During course students will learn to install and setup environment for various technologies.

Learning Outcomes: After completion of this course successfully the students will be able to:

1. Install various Big data technologies.
2. Setup environment for Big data analysis.
3. Perform data analysis on large dataset.

SECTION-A (15L)

1. Big Data and Hadoop: Hadoop architecture, Hadoop Versioning and configuration, Single node & Multi-nodeHadoop, Hadoop commands, Models in Hadoop, Hadoop daemon, Task instance, illustrations.

SECTION-B (15L)

2. Map-Reduce: Framework, Developing Map-Reduce course, Life cycle method, Serialization, Running Map Reducein local and pseudo-distributed mode, illustrations.
HIVE: Installation, data types and commands, illustration.

SECTION-C (15L)

3.SQOOP: Installation, importing data, Exporting data, Running, illustrations

SECTION-D (7+8=15L)

4.PIG: Installation, Schema, Commands, illustrations.

Recommended Books:

1. Hadoop in Action: Chuck Lam, 2010, ISBN: 9781935182191
2. Data- intensive Text Processing with Map Reduce: Jimmy Lin and Chris Dyer, Morgan & Claypool Publishers, 2010

2. Computer Vision

Total Marks : 100		**Max. Time : 3 Hrs.**
End Sem Exam: 60 Marks **Min. Pass Marks : 50%**	**Internal Assessment: 40 Marks** **Min. Pass Marks : 50%**	**Credits:- 3** **Lectures:- 45**

SECTION-A (8+7=15L)

1. Introduction :

Light, Brightness adaption and discrimination, Pixels, coordinate conventions, Imaging Geometry, Perspective Projection, Spatial Domain Filtering, sampling and quantization.

2. Spatial Domain Filtering :

Intensity transformations, contrast stretching, histogram equalization, Correlation and convolution, Smoothingfilters, sharpening filters, gradient and Laplacian.

SECTION-B (8+7=15L)

3. Filtering in the Frequency domain:

Hotelling Transform, Fourier Transforms and properties, FFT (Decimation in Frequency and Decimation in TimeTechniques), Convolution, Correlation, 2-D sampling, Discrete Cosine Transform, Frequency domain filtering.

4. Image Restoration:

Basic Framework, Interactive Restoration, Image deformation and geometric transformations,image morphing, Restoration techniques, Noise characterization, Noise restoration filters,Adaptive filters, Linear, Position invariantdegradations, Estimation of Degradation functions,Restoration from projections.

SECTION-C (6+9=15L)

5. Morphological Image Processing:

Basics, SE, Erosion, Dilation, Opening, Closing, Hit-or-Miss Transform, Boundary,Detection, Hole filling, Connected components, convex hull, thinning, thickening, skeletons,pruning, Geodesic Dilation, Erosion, Reconstruction by dilation and erosion.

6. Image Segmentation:

Boundary detection based techniques, Point, line detection, Edge detection, Edge linking,local processing, regional processing, Hough transform, Thresholding, Iterative thresholding,Otsu's method, Moving averages, Multivariable thresholding, Region-based segmentation,Watershed algorithm, Use of motion in segmentation

SECTION-D (8+7=15L)

Introduction to OpenCV and Python Modules:

setup OpenCV on your computer, Core modules, The Core Functionality, imgproc module: Image Processing,highgui module : High Level GUI and Media, ml module : Machine Learning, video module: Video analysis. Python modules: Pillow, PIL, scikit-image,

Introduction to Medical Image and Processing : What is medical image, file formats, processing, application ofmedical image processing, case studies.

Recommended Books:

1. Richard Szeliski, Computer Vision: Algorithms and Applications .
2. Hartley & Zisserman (HZ) Multiple View Geometry in Computer Vision 2/e
3. Ma, Soatto, Kosecka and Sastry (MaSKS) An Invitation to 3D Vision

3. Natural Language Processing and Text Mining

Total Marks : 100		**Max. Time : 3 Hrs.**
End Sem Exam: 60 Marks	**Internal Assessment: 40 Marks**	**Credits:- 3**
Min. Pass Marks : 50%	**Min. Pass Marks : 50%**	**Lectures:- 45**

Learning Objectives:
The objective of this course is to introduce the fundamentals of Natural Language Processing and Text mining to the students. The aim of this course is to make student understand how to tag a given text with basic Language processing features and design an innovative application using NLP components. This course also aims to introduce the link with NLP and text miningand provides scope to learn text mining to build various applications.

Learning Outcomes: After completion of this course successfully the students will be able to:

1. Process the natural language with ease to build useful applications.
2. Understand the working with Natural Language in computer.
3. Understand the process of text mining
4. Use NLP and text mining to build useful real world applications

SECTION-A (8+7=15L)

Introduction to Natural Language Processing : Natural Language Processing tasks in syntax, semantics, and pragmatics –
Issues - Applications - The role of machine learning - Probability Basics –Information theory – Collocations -N-gramLanguage Models - Estimating parameters and smoothing - Evaluating language models.
Linguistic essentials: Lexical syntax- Morphology and Finite State Transducers - Part of speech Tagging - Rule-

SECTION-B (8+7=15L)

Based Part ofSpeech Tagging - Markov Models - Hidden Markov Models – Transformation based Models - Maximum Entropy Models. Conditional Random Fields
Syntax Parsing: Grammar formalisms and treebanks - Parsing with Context Free Grammars - Features and Unification -Statistical parsing and probabilistic CFGs (PCFGs)-Lexicalized PCFGs.
Semantic Analysis: Representing Meaning – Semantic Analysis - Lexical semantics –Word-sense disambiguation -Supervised – Dictionary based and Unsupervised Approaches - Compositional semantics Semantic Role Labeling and Semantic Parsing – Discourse Analysis.

SECTION-C (15L)

Introduction of Text Mining : Origin of Text Mining - Understanding Text – Applications – Information Visualization - Architecture for Text Mining Applications. Words – Sentences - Indexing Document TextHiddenMarkov Models - POS Taggers - Word Sense disambiguation.

SECTION-D (8+7=15L)

Information Extraction: IE Application - Entity Extraction - IE Systems - Phrase Extraction. Search Engines: Early Search
EnginesIndexing text for Search-Indexing Multimedia – Queries - Searching an index - Viewing search results.
Web Mining: Web Structure - Search Engine Coverage - A distributed Search-Crawlers Visualization Summarization:Training a summarizer - Sentence Selection Information Monitor.

Recommended Books:

1. Daniel Jurafsky and James H. Martin Speech and Language Processing (2nd Edition), Prentice Hall; 2edition, 2008
2. Christopher D. Manning and Hinrich Schuetze , Foundations of Statistical Natural Language Processing by,MIT Press, 1999
3. Steven Bird, Ewan Klein and Edward Loper Natural Language Processing with Python, O'Reilly Media; 1edition, 2009
4. Manu Konchady "Text Mining Application Courseming", Cengage Learning, Fourth Indian Reprint, 2009.
5. Thomas W. Miller, Prentice Hall, "Data and Text Mining - A Business Applications Approach", Secondimpression, 2011

Annexure –I: Syllabus for Elective-II papers

4. HEALTH ANALYTICS

UNIT I

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

Unit II

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

Unit III

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

Unit IV

Advanced Data Analytics Advanced Data Analytics for Healthcare– Review of Clinical Prediction Models- Temporal Data Mining for Healthcare Data- Visual Analytics for Healthcare- Predictive 53 Models for Integrating Clinical and Genomic Data- Information Retrieval for Healthcare- Data Publishing Methods in Healthcare.

Unit V

Applications Applications and Practical Systems for Healthcare– Data Analytics for Pervasive HealthFraud 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

- Chandan K. Reddy and Charu C Aggarwal, "Healthcare data analytics", Taylor & Francis, 2015.

REFERENCE BOOKS

- Hui Yang and Eva K. Lee, "Healthcare Analytics: From Data to Knowledge to Healthcare Improvement, Wiley, 2016.

5. Time Series Analysis and Forecasting

Total Marks : 100		**Max. Time : 3 Hrs.**
End Sem Exam: 60 Marks **Min. Pass Marks : 50%**	**Internal Assessment: 40 Marks** **Min. Pass Marks : 50%**	**Credits:- 3** **Lectures:- 45**

Learning Objectives:
This course is focus towards analysis of time series data and make prediction i.e. forecasting based on the outcome. This typeof analysis is very useful in business and finance.
Learning Outcomes: After completion of this course successfully the students will be able to:

1. Understand the inherit difference with normal data and time series data.
2. Perform various analysis on time series data.
3. Derive conclusion from time series data.

SECTION-A (15L)

1. Basics of Time series : A model Building strategy, Time series and Stochastic process, stationarity, Auto correlation,
meaning and definition – causes of auto correlation - consequence of autocorrelation – test for auto – correlation. Study ofTime Series model and their properties using correlogram, ACF and PACF. Yule walker equations

SECTION-B (15L)

2. Time Series Models : White noise Process, Random walk, MA, AR, ARMA and ARIMA models, Box-Jenkins's Methodology fitting of AR(1), AR(2), MA(1), MA(2) and ARIMA(1,1) process. Unit root hypothesis, Co-integration,Dicky Fuller test unit root test, augmented Dickey – Fuller test.

SECTION-C (15L)

3. Non-linear time series models :ARCH and GARCH Process, order identification, estimation and diagnostic tests and forecasting. Study of ARCH (1) properties. GARCH (Conception only) process for modelling volatility.

SECTION-D (15L)

4. Multivariate Liner Time series: Introduction, Cross covariance and correlation matrices, testing of zero cross correlation and model representation. Basic idea of Stationary vector Autoregressive Time Series with order one: ModelStructure, Granger Causality, stationary condition, Estimation, Model checking.

Recommended Books:

1. Box, G. E. P. and Jenkins, G. M. (1976). Time Series Analysis – Forecasting and Control, Holden – day, SanFrancisco.
2. Chatfield, C. (2003) Analysis of Time Series, an Introduction, CRC Press.
3. Ruey S. Tsay (2005). Analysis of Financial Time Series, Second Ed. Wiley & Sons.
4. Ruey S. Tsay (2014). Multivariate Time series Analysis: with R and Financial Application, Wiley & Sons.
5. Introduction to Statistical Time Series: W.A. Fuller

6. Product Development

Total Marks : 100		**Max. Time : 3 Hrs.**
End Sem Exam: 60 Marks **Min. Pass Marks : 50%**	**Internal Assessment: 40 Marks** **Min. Pass Marks : 50%**	**Credits:- 3** **Lectures:- 45**

Learning Objectives:
The objective of this course is to train students on software development life cycle and familiarize them with various stage ofdevelopment by giving example through various case studies.
Learning Outcomes: After completion of this course successfully the students will be able to:

1. Sketch plans for product development
2. Develop server side logic and expose them as API.
3. Build web client using web Technologies (HTML, CSS and JS).
4. Build mobile client using Android.
5. Implement their idea into product.

SECTION-A (15L)

1. Software Development Life Cycle

2. Building web client / Front End :

HTML5, CSS, Bootstrap, Flex, JavaScript, (React, vue.js etc.)

SECTION-B (5+10=15L)

3. Building Android client :

Android Studio setup and Basic of Android Programming, how to use various API

SECTION-C (15L)

4. Server Side Technologies (Flask and Django) :

System setup of Flask and Django, learning basic of Flask and Django, Learning REST API development, Testing API usingPostman.

SECTION-D (5+10=15L)

5. Deployment and Maintenance :
Cloud-based and individual web server deployment, Continuation deployment and integration (Travis CI, GitLab CI)

6. Minimum two case studies: One web application and one Android application. Must use Machine learning and Datascience.

Recommended Books:

1. Willi Richert, Luis Pedro Coelho , Building Machine Learning Systems with Python, Packt Publishing Limited.
2. John Horton, Android Programming for Beginners, Packt Publishing Limited.

Intake-120

Eligibility: Any 12th/MCVC/Diploma pass Student