**Savitribai Phule Pune University**
**(Formerly University of Pune)**

**Department of Technology**

# STRUCTURE OF FOUR-YEAR FULL-TIME DEGREE PROGRAM IN

## B.Sc. Data Science

**(Session 2023-24)**

---

**Eligibility: Any 12th Passed**

**Course Intake: 240**

**After Completion of 6 Semesters Candidate will be awarded with BSc Data science Degree and after Completion of 8th semester, which is optional Candidate, will be awarded with BSc Data Science Honor.**

| Semester | Lectures per Week | Course Credits |
|---|---|---|
| 1st Semester | 22 | 22 |
| 2nd Semester | 22 | 22 |
| 3rd Semester | 22 | 22 |
| 4th Semester | 22 | 22 |
| 5th Semester | 22 | 22 |
| 6th Semester | 22 | 22 |
| 7th Semester | 23 | 22 |
| 8th Semester | 23 | 22 |
| Total | | Course Credits – 176 |

**INDEX**



***In Seminar-1, Seminar-2, Seminar-3 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 |
| BSCDS1 | Python Programming | 2 | | 2 | 2 | | 2 | 4 |
| BSCDS2 | Applied Mathematics | 2 | | 2 | 2 | | 2 | 4 |
| BSCDS3 | Data science with R | 2 | | 2 | 2 | | 2 | 4 |
| BSCDS4 | Python Programming Lab | | 2X2=4 | | | 4 | | 4 |
| BSCDS5 | Data science with R Lab | | 2X2=4 | | | 4 | | 4 |
| BSCDS6 | English Communication-I | 2 | | | 2 | | | 2 |
| **Total** | | 12 | 4 | 4 | 14 | 4 | 4 | 22 |
| **Total Contact Hours per Week=22** | | | | | | | **Total Credits=22** | |

**Semester-II**

| Subject Code | Subjects Name | Contact Hours per Week | | | Credits | | | |
|---|---|---|---|---|---|---|---|---|
| | | L | P | T | L | P | T | Total |
| BSCDS7 | Probability and Statistics | 2 | | 2 | 2 | | 2 | 4 |
| BSCDS8 | Web Framework | | 2X2=4 | | | 4 | | 4 |
| BSCDS9 | Data Science with Python | 2 | | 2 | 2 | | 2 | 4 |
| BSCDS10 | Operating System | 2 | | | 2 | | | 2 |
| BSCDS11 | Data Analytic and Visualization | 2 | | 2 | 2 | | 2 | 2 |
| BSCDS12 | Data Analytic andVisualization Lab | | 2X1=2 | | | 2X=2 | | 2 |
| BSCDS13 | English Communication-II | 2 | | 2 | 2 | | 2 | 4 |
| **Total** | | 8 | 6 | 8 | 8 | 6 | 8 | 22 |
| **Total Contact Hours per Week=22** | | | | | | | **Total Credits=22** | |

**B.Sc. Data Science**
**Semester-III**

| Subject Code | Subjects Name | Contact Hours perWeek | | | Credits | | | |
|---|---|---|---|---|---|---|---|---|
| | | L | P | T | L | P | T | Total |
| BSCDS14 | Optimization Techniques | 2 | | 2 | 2 | | 2 | 4 |
| BSCDS15 | Database Management System(DBMS) | 2 | | 2 | 2 | | 2 | 4 |
| BSCDS16 | Machine Learning | 2 | | 2 | 2 | | 2 | 4 |
| BSCDS17 | Optimization Techniques(Python) -LAB | | 2X1=2 | | | 2 | | 2 |
| BSCDS18 | DBMS - LAB | | 2X1=2 | | | 2 | | 2 |
| BSCDS19 | Machine Learning - Lab | | 2X2=4 | | | 4 | | 4 |
| BSCDS20 | Seminar-1 | | | 2 | | | 2 | 2 |
| **Total** | | **6** | **8** | **8** | **6** | **8** | **8** | **22** |
| **Total Contact Hours per Week=22** | | | | | | | **Total Credits=22** | |

**Semester-IV**

| Subject Code | Subjects Name | Contact Hours per Week | | | Credits | | | |
|---|---|---|---|---|---|---|---|---|
| | | L | P | T | L | P | T | Total |
| BSCDS21 | Deep Learning | 2 | | 2 | 2 | | 2 | 4 |
| BSCDS22 | IoT Programming and BigData | 2 | | 2 | 2 | | 2 | 4 |
| BSCDS23 | Data warehouse and Data Mining | 2 | | 2 | 2 | | 2 | 4 |
| BSCDS24 | Deep Learning - LAB | | 2X2=4 | | | 4 | | 4 |
| BSCDS25 | IoT Programming and BigData LAB | | 2X1=2 | | | 2 | | 2 |
| BSCDS26 | Data warehouse and Data Mining LAB | | 2X1=2 | | | 2 | | 2 |
| BSCDS27 | Seminar-2 | | | 2 | | | 2 | 2 |
| **Total** | | **6** | **8** | **8** | **6** | **8** | **8** | **22** |
| **Total Contact Hours per Week=22** | | | | | | | **Total Credits=22** | |

**B.Sc. Data Science**
**Semester-V**

| Subject Code | Subjects Name | Contact Hoursper Week | | | Credits | | | |
|---|---|---|---|---|---|---|---|---|
| | | L | P | T | L | P | T | Total |
| BSCDS28 | Big Data Analytics through Spark | 2 | | 2 | 2 | | 2 | 4 |
| BSCDS29 | Introduction to ArtificialIntelligence | 2 | | 2 | 2 | | 2 | 4 |
| BSCDS30 | Machine Learning Operations(ML Ops) | 2 | | 2 | 2 | | 2 | 4 |
| BSCDS31 | Elective – I | 2 | | 2 | 2 | | 2 | 4 |
| BSCDS32 | Big Data Analytics through Spark-LAB | | 2X1=2 | | | 2 | | 2 |
| BSCDS33 | Artificial Intelligence (PROLOG / Python) - LAB | | 2X1=2 | | | 2 | | 2 |
| BSCDS34 | Project Work –Minor (IoT/Machine Learning) | | 2X1= 2 | | | 2 | | 2 |
| **Total** | | **8** | **6** | **8** | **8** | **6** | **8** | **22** |
| **Total Contact Hours per Week=22** | | | | | | | **Total Credits=22** | |

**Semester-VI**

| Subject Code | Subjects Name | Contact Hours per Week | | | Credits | | | |
|---|---|---|---|---|---|---|---|---|
| | | L | P | T | L | P | T | Total |
| BSCDS35 | NoSQL Databases | 2 | | 2 | 2 | | 2 | 4 |
| BSCDS36 | Cloud Computing | 2 | | 2 | 2 | | 2 | 4 |
| BSCDS37 | Big Data Acquisition And Analysis | 2 | | 2 | 2 | | 2 | 4 |
| BSCDS38 | Elective - II | 2 | | 2 | 2 | | 2 | 4 |
| BSCDS39 | Big Data Acquisition and Analysis Lab | | 2X1=2 | | | 2 | | 2 |
| BSCDS40 | NoSQL Databases-LAB | | 2X1=2 | | | 2 | | 2 |
| BSCDS41 | Project Work – Major | | 2X1=2 | | | 2 | | 2 |
| **Total** | | **8** | **6** | **8** | **8** | **6** | **8** | **22** |
| **Total Contact Hours per Week=22** | | | | | | | **Total Credits=22** | |

**B.Sc. Data Science**
**Semester-VII**

| Subject Code | Subjects Name | Contact Hours per Week | | | Credits | | | |
|---|---|---|---|---|---|---|---|---|
| | | L | P | T | L | P | T | Total |
| BSCDS42 | Generative AI - I | 2 | | 2 | 2 | | 2 | 4 |
| BSCDS43 | Reinforcement Learning | 2 | | 2 | 2 | | 2 | 4 |
| BSCDS44 | Research Methodology | 2 | | 2 | 2 | | 2 | 4 |
| BSCDS45 | Generative AI -I LAB | | 2X1=2 | | | 2 | | 2 |
| BSCDS46 | Reinforcement Learning-LAB | | 2X1=2 | | | 2 | | 2 |
| BSCDS47 | Capstone Project-I | | 2X2=4 | | | 4 | | 4 |
| BSCDS48 | Massive Open Online Courses(MOOCs) | | | 2 | | | 2 | 2 |
| **Total** | | **6** | **8** | **8** | **6** | **8** | **6** | **22** |
| **Total Contact Hours per Week=22** | | | | | | | **Total Credits=22** | |

**B.Sc. Data Science**
**Semester-VIII**

| Subject Code | Subjects Name | Contact Hours per Week | | | Credits | | | |
|---|---|---|---|---|---|---|---|---|
| | | L | P | T | L | P | T | Total |
| BSCDS49 | Generative AI - II | 2 | | 2 | 2 | | 2 | 4 |
| BSCDS50 | Data Engineering | 2 | | 2 | 2 | | 2 | 4 |
| BSCDS51 | Data Visualization-Tableau | 2 | | 2 | 2 | | 2 | 4 |
| BSCDS52 | Generative AI – II LAB | | 2X1=2 | | | 2 | | 2 |
| BSCDS53 | Data Engineering -LAB | | 2X1=2 | | | 2 | | 2 |
| BSCDS54 | Capstone Project-II | | 2X2=4 | | | 4 | | 4 |
| BSCDS55 | Seminar – 3 | | | 2 | | | 2 | 2 |
| **Total** | | **6** | **8** | **8** | **6** | **8** | **6** | **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**

**BSCDS1- Python Programming**

| | | |
|---|---|---|
| **Total Marks: 100** | | **Max. Time: 3 Hrs.** |
| **End Sem Exam: 100 Marks**<br>**Min. Pass Marks: 40%** | **Internal Assessment: 50 Marks**<br>**Min. Pass Marks: 40%** | **Credits: - 4**<br>**Lectures: - 42** |

**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 datastructures.

**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.

**Python Programming:** 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, Flowchartsand Pseudo-codes, implementation of algorithms. Basics of Python Programming: Features ofPython, 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. Otherdata types-Tuples, Lists and Dictionary. Need for functions, Function: definition, call, variablescope and lifetime, the return statement. Defining functions, Lambda or anonymous function, documentation string, good programming practices. Introduction to Units, Introduction to packages in Python, Introduction to standard library Units.

**Strings Linear Data Structures:** 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 stringUnit. 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.

**Object Oriented Programming & File Handling**:- 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).

**Linked List, Stacks:-**Introduction to Static and Dynamic Memory Allocation, Linked List: Introduction, of LinkedLists, 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 LinkedList, Doubly Circular Linked List. Basic concept, stack Abstract Data Type, Representation ofStacks Using Sequential Organization, stack operations, Multiple Stacks, Applications of Stack- Expression Evaluation and Conversion, Polish notation and expression conversion.

**Queue :-**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).

**Working with Data in Python**: -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: 9780132492642, ISBN-13: 978- 0132492642
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

10.Reema Thareja, “Python Programming Using Problem Solving Approach”, Oxford University Press,ISBN 13: 978-0-19-948017-6

11.R. Nageswara Rao, “Core Python Programming”, Dreamtech Press; Second edition ISBN10:938605230X, ISBN-13: 978-9386052308 ASIN: B07BFSR3LL

**BSCDS2- Applied Mathematics**

| | | |
|---|---|---|
| **Total Marks: 100** | | **Max. Time: 3 Hrs.** |
| **End Sem Exam: 100 Marks** | **Internal Assessment: 50 Marks** | **Credits: - 4** |
| **Min. Pass Marks: 40%** | **Min. Pass Marks: 40%** | **Lectures: - 42** |

**Learning Objectives:**

The course is a brief overview of the basic tools from Linear Algebra and Multivariable Calculus that will beneeded 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 MultivariableCalculus 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.

**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 andnon-singular matrices, examples, finding an inverse matrix. Introduction**:** Sets - finite and infinite sets, Infinite Sets; functions, relations, Properties of Binary Relations, Closure, Partial Ordering Relations; counting - Pigeonhole Principle, Permutation andCombination; Mathematical Induction, Principle of Inclusion and Exclusion.

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, identity matrix,triangular matrix, idea about sparse and dense matrix, unit vectors, symmetric matrix, Hermitian, skew- Hermitian and unitary matrices.

**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 coefficients and their solution, Substitution Method, Recurrence Trees, Master Theorem.

**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.

**Real-valued functions of two or more variables**:-Definition, examples, simple demos, applications.

Prepositional Logic: Logical Connectives, Well-formed Formulas, Tautologies, Equivalences, Inference Theory.Analysis elements Distance, Limits, Continuity, Differentiability, the gradient and the Gaussian. Optimizationproblems Simple examples, motivation, the role of the Hessian maxima and minima and related extreme conditions. Integration Double integrals, Fubini's theorem, properties, applications.

Numerical Linear Algebra**:** Regularization, Principal Component Analysis, Singular-Value, Decomposition,Latent Semantic Analysis,Case Studies: Recommender Systems, Page Ranking

**References:**


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.

## BSCDS3- DATA SCIENCE WITH R

| | | |
|---|---|---|
| **Total Marks :100** | | **Max. Time: 3 Hrs.** |
| **End Sem Exam: 100 Marks**<br>**Min. Pass Marks: 40%** | **Internal Assessment: 50 Marks**<br>**Min. Pass Marks: 40%** | **Credits: - 4**<br>**Lectures: - 42** |

**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.

**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.

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 & Sub setting.

**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.

**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.

**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, detailsof the first node, xml to data node.

**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 UniversityPress, 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

**BSCDS4- Python Programming Lab**

| | | |
|---|---|---|
| **Total Marks: 100** | | **Max. Time: 3 Hrs.** |
| **End Sem Exam: 100 MarksMin. Pass Marks: 40%** | **Internal Assessment: 50 Marks Min. Pass Marks: 40%** | **Credits: - 4** |
| | | **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 datastructures.

**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.

**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

## BSCDS5- R PROGRAMMING LAB

| | | |
|---|---|---|
| **Total Marks: 100** | | **Max. Time: 3 Hrs.** |
| **End Sem Exam: 100 Marks** | **Internal Assessment: 50 Marks** | **Credits: - 4** |
| **Min. Pass Marks: 40%** | **Min. Pass Marks: 40%** | **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

## BSCDS6-English Communication-I

| | | |
|---|---|---|
| **Total Marks: 100** | | **Max. Time: 3 Hrs.** |
| **End Sem Exam: 100 Marks**<br>**Min. Pass Marks: 40%** | **Internal Assessment: 50 Marks**<br>**Min. Pass Marks: 40%** | **Credits: - 4**<br>**Lectures: - 30** |

**Learning Objectives:**

The purpose of this course is twofold: to train students in communication skills and to help develop in them a facilityfor communicative English. Since language, it is which binds society together and serves as a crucial medium of interaction as well as interchange of ideas and thoughts, it is important that students develop a capacity for clear andeffective 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 inprofessional spheres where English is the indispensable tool of communication.

**Introduction:-**What is communication, Types of communication: Horizontal, Vertical, Interpersonal, Grapevine Uses of Communication,

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

http://tsime.uz.ac.zw/claroline/backends/download.php?url=L0ludHJvX3RvX2NvbW11bmljYXRpb25fVGhlb3J 5LnBkZg%3 D%3D&cidReset=true&cidReq=MBA563

**Language of Communication:-**Verbal: spoken and written**,** Non-verbal: Proxemics, Kinesics, Haptics, Chronemics, Paralinguistic**,** Barriers to communication.

**Reading Comprehension**:-Locate and remember the most important points in the reading, Interpret and evaluate events, ideas, and information, Read "between the lines" to understand underlying meanings, Connect information to what they already know.

**Writing:**

1. Expanding an Idea
2. Note Making
3. Memo
4. Writing Formal Email

5. Writing a Business Letter
6. Report Writing
7. Building Resume
8. Video Resume

**B.Sc. Data Science**

**Semester-II**

**BSCDS7- Probability and Statistics**

| | | |
|---|---|---|
| **Total Marks: 100** | | **Max. Time: 3 Hrs.** |
| **End Sem Exam: 100 Marks** | **Internal Assessment: 50 Marks** | **Credits: - 4** |
| **Min. Pass Marks: 40%** | **Min. Pass Marks: 40%** | **Lectures: - 42** |

**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.

**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 - FundamentalPrinciple 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.

**Random Variables and Probability Distributions** - Random Variables - Discrete Probability Distributions - DistributionFunctions 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 GeometricProbability. Measures of central tendency: **–** properties – merits and demerits – weighted means– graphical location ofmedian, 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.

**Mathematical Expectation** - Definition of Mathematical Expectation - Functions of Random Variables - Theorems on Expectation - Variance & Standard Deviation - Theorems on Variance - Standardized Random Variables - Special ProbabilityDistributions - Binomial Distribution - Normal Distribution - Poisson distribution. Skewness and Kurtosis – Pearson's coefficient of skewness – Bowler's coefficient of skewness –coefficient of skewness 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 twoindependent means, Inference for comparing two paired means, Power, Comparing more than two means, ANOVA, Conditions for ANOVA, Multiple comparisons, Bootstrapping.

**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 CoefficientGeneralized Correlation Coefficient 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-SquareGOF Test, The Chi-Square Independence Test.

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

**References**:

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.

**BSCDS8- Web Framework**

| | | |
|---|---|---|
| **Total Marks: 100** | | **Max. Time: 3 Hrs.** |
| **End Sem Exam: 100 Marks**<br>**Min. Pass Marks: 40%** | **Internal Assessment: 50 Marks**<br>**Min. Pass Marks: 40%** | **Credits: - 4**<br>**Lectures: - 42** |

**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:

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

Introduction to Django:Understanding web development frameworks,Introduction to Django and its features,Installing Django and setting up a development environment,Creating a simple Django project and app

**Django Models and Database Integration:**Creating models and defining database tables,Working with Django's Object-Relational Mapping (ORM),Performing database queries using Django's QuerySet API,Qw2003 Migrations and database schema evolution

**Views and Templates**: Building views to handle HTTP requests,Creating templates for dynamic HTML generation,Routing and URL patterns in Django.Passing data from views to templates
Django Forms:Creating HTML forms in Django,Form validation and handling form submissions,Customizing form behavior with Django form classes,Integrating forms with models
Django Admin Panel:Utilizing the Django admin interface for content management,Customizing the admin panel for specific models,Adding custom actions and filters

**Authentication and Authorization**: Implementing user authentication in Django, Managing user sessions and passwords Configuring permissions and authorization

**Django REST Framework**: Introduction to RESTful APIs, Building APIs with Django REST Framework,Serializers, views, and authentication for APIs, Consuming APIs in Django applications

**Frontend Integration with Django**: Integrating frontend frameworks (e.g., Bootstrap) with Django, Using static files and media in Django projects,AJAX and asynchronous behavior in Django applications

**Testing and Debugging in Django**: Writing unit tests for Django applications, Debugging techniques and tools,Best practices for testing in Django
Deployment and Scaling: Preparing a Django application for deployment, Choosing a hosting platform (e.g., Heroku, AWS),Configuring production settings, Scaling Django applications

**Advanced Topics**: Signals and event handling in Django, Building custom middleware, Caching strategies in Django, Internationalization and localization

**Project Work**: Applying knowledge to a real-world project, Working on a comprehensive Django project from start to finish, Code reviews and best practices

**BSCDS9- Data Science with Python**

| | | |
|---|---|---|
| **Total Marks: 100** | | **Max. Time: 3 Hrs.** |
| **End Sem Exam: 100 Marks** | **Internal Assessment: 50 Marks** | **Credits: - 4** |
| **Min. Pass Marks: 40%** | **Min. Pass Marks: 40%** | **Lectures: - 42** |

**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 datastructures.

**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.

**Numpy and Pandas Packages** NumPy and array - 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 Data Frame data structures in pandas - Creation of Data Frames – Accessing the columns in a Data Frame - Accessing the rows in a Data Frame - Panda's Index Objects - Reindexing Series and Data Frames - 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.

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

**Data Wrangling** Combining and Merging Data Sets – Reshaping and Pivoting – Data Transformation – StringManipulations – Regular Expressions.

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

## BSCDS10- Operating System and Information Security

| | | |
|---|---|---|
| **Total Marks: 100** | | **Max. Time: 3 Hrs.** |
| **End Sem Exam: 100 Marks** | **Internal Assessment: 50 Marks** | **Credits: - 2** |
| **Min. Pass Marks: 40%** | **Min. Pass Marks: 40%** | **Lectures: -42** |

**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.

**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. Operating System Organization Processor and user modes, kernels, system calls and systemcourses. 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 inter process communication; deadlocks. Memory Management: Physical and virtual address space; memory allocation strategies –fixed andvariable partitions, paging, segmentation, virtual memory. File and I/O Management: Directory structure, file operations, files allocation methods, device management. Protection and Security: Policy mechanism, Authentication, Internal access Authorization**.**

**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, No malicious Program Errors, viruses and other malicious code, TargetedMalicious code, controls Against Program Threats, Protection in General- Purpose operating system protected objects and methods of protection memory and admen's 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

**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.

**References:**


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


## BSCDS11- Data Analytic and Visualization

| | | |
|---|---|---|
| **Total Marks: 100** | | **Max. Time: 3 Hrs.** |
| **End Sem Exam: 100 MarksMin. Pass Marks: 40%** | **Internal Assessment: 50 Marks Min. Pass Marks: 40%** | **Credits: - 2** |
| | | **Lectures: - 42** |

**Learning Objectives:**

1. Make students understand about commonly used terms and techniques related to data analytics that are in use today.
2. To discuss and train students on how managers to make better decisions can use data and data analytics.
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 1** : **Foundation of Data Analytics & Python Programming**:

Foundation of Data Analytics: - Introduction ,Evolution , Concept and Scopes , Data , Big Data, Metrics and Data classification, Data Reliability & Validity, Problem Solving with Analytics, Different phases of Analytics in the business and Data science domain, Descriptive Analytics, Predictive Analytics and Prescriptive Analytics , Different Applications of Analytics in Business, Text Analytics and Web Analytics, Skills for Business Analytics , Concepts of Data Science, Basic skills required for understanding Data Science.

**Section 2**. : **Introduction to SQL, PostgreSQL and Business Intelligence**:

Learning SQL query (DDL,DML,DCL) structure with examples, Data management and query system OLTP and OLAP and Their data models, Data warehousing, ET Concepts of Business intelligence (BI), the relevance of BI in application to analytics industry and different domains.

**Section 3**. **Introduction to Power BI:-**

Power BI is the newest Microsoft Business Intelligence and Data Analysis tool. In this module, we will go through the basics of this product and introduce all components of Power BI (Power Query, Power Pivot, Power View, and Power Q&A). You will see some demos and introductions about Power BI desktop, Office 365 Power BI subscription, and Power BI website, and mobile apps. You will see some basic demos of how easy it to use is Power BI in some scenarios.

- Introduction to Power BI: What is Power BI?
- Power BI Desktop; The First Experience
- Power BI Website; You'll Need Just a Web Browser
- Introduction to Power BI Components: Power Query, Power Pivot, Power View, and Power Q&A.

## References:


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

**BSCDS12- Data Analytic and Visualization Lab**

| | | |
|---|---|---|
| **Total Marks: 50** | | **Max. Time: 2 Hrs.** |
| **End Sem Exam: 50 MarksMin. Pass Marks: 40%** | **Internal Assessment: 50 Marks Min. Pass Marks: 40%** | **Credits: - 2** |
| | | **Lectures: - 42** |

1 : **Getting Data and Transformation**

Getting Data is the first experience of working with Power BI. You can connect to many data sources on-premises or on the cloud. In this section, you will learn how the Get data experience in Power BI is and how you can transform the data in a way to get it ready for modelling.

- What is Power Query: Introduction to Data Mash-Up Engine of Power BI
- Different versions of Power Query
- Power Query Introduction
- Query Editor
- Transformation GUI
- Get Started with Power Query: Movies Data Mash-Up
- Power BI Get Data from Excel: Everything You Need to Know
- What is the Role of Power Query in a Power BI Solution

2 : **Data Modelling and Simple DAX**:

Data Modelling in Power BI is an in-memory-based technology. You will learn about the structure of modeling in Power BI, and you will learn the importance of relationships and their direction. You will also learn about calculations in Power BI and how to write them. DAX is the Data Analytical eXpression language. DAX has a similar structure to excel functions, but it is different. In this module, you will learn DAX about DAX too.
The content that you will learn in this module includes but is not limited to;

- Power BI xVelocity engine basics and concepts
- Relationships in Power BI
- Hierarchies and Formatting
- Sorting by other columns
- Date Table
- Introduction to DAX
- Calculated Columns, Measures, and Calculated Tables

3. **Data Modelling with Power BI**:-

Data modelling concepts such as the best practice of designing data tables and how to build a data model that performs fast, building star-schema, and fact and dimension tables are all explained in this training.
At the end of this training, you will learn how to build a star schema from a data source that covers the requirement. You will understand the concept of Power BI relationships and can build a data model based on best practices.

This training includes but is not limited to the topics below;

4 : **Relationships:**

Understanding the relationships is one of the most essentials learning to build a data model. This training starts with an explanation of why relationships are needed, what are different types of relationships, and attributes such as the direction or cardinality of the relationship are covered with examples.

- Why relationships in Power BI
- one-to-many, many-to-one, many-to-many, and one-to-one relationships
- What is the direction of the relationship?
- role-playing dimension and inactive relationships
- relationships based on multiple columns

4.1: **Dimensional modelling:-**

In this section, you will learn what is star-schema, and what are fact and dimension tables. What are different types of fact tables, and why the star-schema design is important for data modeling? You will learn the principles of data modeling.

- Why data preparation
- What is dimension table?
- What is fact table and different types of it?
- Do you need a date dimension?
- Power BI default or custom date table
- What is star-schema?

4.2: **Star schema in action**

After learning the concepts of star-schema, it is time to put that in action and learn how you can build a star-schema model in Power BI using techniques of data transformation and the concepts of modeling.

- Combining dimension tables
- Creating shared dimension
- Combine tables or create relationships?
- What fields to hide?
- Build your first star-schema
- One dimension filters another dimension

4.3: **Better data model**

Finally, there are important tips to consider to take your data model to the next level, which is covered in this section of the training.

- Move shared tables to dataflows
- Shared datasets and how to use it
- Reducing the size of the model
- Important consideration about dates in Power BI

5. **Power Query: Get Data and Transform**:

You will learn about different types of transformations available in Query Editor. Table transformations such as Pivot and Unpivot will be discussed, as well as specific column transformations such as date column transformations. You will also learn about M (Power Query Formula Language). Unique features such as error handling, generators, structured columns, custom functions, and many other advanced level features of Power BI data transformations will be explored through lectures.

After this course, you will be able to implement any type of data transformation through Power Query in Excel or Power BI. You will be able to work through your raw data and make it ready for modelling and analytics.

The training includes but is not limited to the topics below:

5.1: **Get Data**

In this section, you will learn about Power Query basics which start with getting data. You will learn that Power Query is the data transformation tool in Power BI. You will learn different parts of the Query Editor through an example of using Power Query to transform a dataset.

- Introduction to Power Query
- Query Editor
- Get Data from Web

- Basic Transformations
- Get Data from Excel
- Use First Row As Headers / Use Headers as First Row
- Get Data from SQL Server

5.2: **Data Types and Data Structures**

Before going any further in learning Power Query, you need to understand data structures and data types. There are three main data structures in Power Query; table, record, and list. You will learn about these types through an example of getting data from a JSON structure. You will also learn about data types and their differences.

- Base data structures in Power Query
- Get Data from JSON
- Transforming Table, Record, and List
- Data Types in Power Query
- Query Operations
- Enable Load; Performance Boost
- Query Operations; Duplicate, and Reference

5.3: **Combine Queries**:-

One of the most common data transformations is combining datasets. Depends on the types of datasets and the way that they are related to each other, you may want to merge or append them. In this section, you will learn why you need to combine data at first, and then you will learn about scenarios that you combine data in Power Query.

- Dimensional Modelling; Designing the data model
- Append, creating a single big query of the same structure
- Merge; joining queries when the structure is different
- Join types in Merge
- Tips to consider after Merge or Append

5.4: **Better Power Query Editor Experience**:-

To get the best experience with Power Query Editor, you need to consider organizing your queries and steps in the right way. In this section, you will learn about actions you can do on steps, such as moving them up or down, splitting steps in a query, etc. You will also learn about organizing your queries in groups (folders).

- Groups; Folders in Query Editor
- Steps Operations
- Splitting query steps
- Moving steps up or down
- Add as new query / Drill Down
- Be Careful of Actions; Undo!

5.5: **Reducing Number of Rows; Filtering**

Filtering rows in Power Query is an important transformation especially when the dataset is big, or when the data needs to be cleaned. There are different ways of doing filtering in Power Query. You will learn about ways to remove some rows from the top or bottom of the table, and ways that you can filter a data table based on criteria. You will learn about basic filtering and the difference between that with advanced filtering, and potential challenges that you may have through this process.

- Row Operations; removing rows
- Row Operations; Keeping rows
- Remove/Keep Errors
- Remove/Keep Duplicates
- Using Remove/Keep combination for troubleshooting report
- Filtering based on Individual values
- Advanced Filtering
- Sorting

5.6: **Column Operations**:-

A data table in Power Query can get big if you don't care about columns. In this section, you will learn actions that you can do on columns, and what are best practices to make sure you have the best performance in your Power BI model considering columns in your tables. You will also learn about some generic column operations and transformations.

- Column Operations
- Choosing Columns
- Removing Columns

- Data Type Change
- Locale consideration for the data type
- Replace Values
- Fill Down/Up; Very Useful for Excel

5.7: **Table Transformations**:-

Some of the most important table transformations will be explained in this section. You will learn about a way to change the granularity of a table; Grouping. You will also learn scenarios that in grouping data can be more than a simple transformation. You will learn about transformations such as Transpose, Pivot, and Unpivot, and the difference of all these items with scenarios of using them on real-world datasets.

- Group By; changing the granularity of the data table
- Group by Advanced
- Scripting and Group by; First and Last item in each group
- Transpose; rows to columns and reverse
- Pivot; changing the name-value structure to columns
- Unpivot; changing the budget column structure to rows

5.8: **Text Transformations**:-

When you work with text values, there are many transformations you can apply. Transformations such as a split column, removing part of a text, or adding a prefix or postfix to it, concatenating some columns together, etc.

- Split Column by Delimiter
- Split Column by number of Characters
- Split into rows instead of columns
- Merge (Concatenate)
- Format
- The difference between Clean and Trim

5.9: **Numeric Transformations**

You will learn in this section how to do numeric transformations. We will talk about standard transformations such as divide and integer-divide. You will also learn about transformations such as rounding, statistics transformations, and even some scientific transformations and use cases for those items.

- Standard transformations; Divide, Integer-divide, Multiply, Add etc.
- Scientific transformations; logarithm, power square, etc.
- Statistics transformations;
- Rounding
- Information functions; Is Even, Is Odd, and Sign.
- Dealing with faults in Numeric calculations

5.10: **Add Column Transformations**

There are two types of transformations in Power Query; Transforming an existing column, or adding a column based on a transformation. In this section, you will learn about these two types, their differences, and few other transformations that we have available in the add column tab of the Power Query Editor through some examples.

- Add Column vs. Transform?
- Add Column with a Transformation
- Index Column: Row Number
- Conditional Column
- Add Column by Example; When you don't know which transformation to use
- Add Custom Column: Generic

6. **Creating your First Business Intelligence Project**: Creating Data source, Creating Data view. Modifying the Data view. Creating Dimensions, Time, and Modifying dimensions. Parent-Child Dimension. 1.4: Data Visualization

6.1 **Data Visualization is the front end of any BI application**; this is the user viewpoint of your system. It is critical to visualize measures, and dimensions effectively so the BI system could tell the story of the data. In this module, you will learn conceptual best practices of data visualizations which are valid through all data visualization tools. You will learn Power BI visualization skills. You will learn how to create effective charts, and dashboards using these tools as well as best practices for working with Power BI Desktop.

The contents are as follows:-

- Power BI Desktop Visualization
- Formatting Visuals in Power BI Desktop
- Basic Charts and Visuals in Power BI

- Sorting, Filtering, and categorization
- Custom Visuals in Power BI Desktop
- Maps and Geo-Spatial Visualization

6.2 **Best Practice Scenarios of Using Visuals**

It is a good time to learn about all built-in visuals in Power BI now. It is important to know which visual should be used in what scenario. You will also learn in this section about the pros and cons of each visual. You will learn specific features of the visual that can help to solve real-world scenarios. All the examples will be hands-on.

- Line Chart; Trend Analysis
- Combo Charts; Different Scales in one graph
- Ribbon Chart; Ranking
- Waterfall Chart; Cashflow
- Scatter Chart; Storytelling with the data
- Grouping charts: Pie, Donut chart, and Treemap charts
- KPI visual
- Gauge
- Card Visual
- Table and Matrix; showing the numbers with conditional formatting

## BSCDS13-English Communication II

| | | |
|---|---|---|
| **Total Marks: 100** | | **Max. Time: 3 Hrs.** |
| **End Sem Exam: 100 MarksMin. Pass Marks: 40%** | **Internal Assessment: 50 Marks Min. Pass Marks: 40%** | **Credits: - 4** |
| | | **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 well as 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 inprofessional spheres where English is the indispensable tool of communication. Overcoming Communication challenges.

**Introduction to English Communication: -** Importance of effective Communication. The role of communication in science and technology, types of communication (verbal and Non-verbal), Overcoming Communication Challenges.

**Speaking and Presentation skills:** Developing Confidence in speaking, Public speaking and Presentation techniques, voice Modulation and Articulation, Non-verbal Communication (Body language, gestures),Effective use of Visual Aids(slides, Charts, graph)

**Listening and Comprehension:-**Active listening skill, note-taking techniques, Understanding different accents, listening to technical and scientific Discourse, Summarizing and paraphrasing information.

**Reading and Comprehension:** Developing Reading habits, Reading Strategies for technical text, Skimming and Scanning, Critical reading and Analyzing Scientific Articles, Vocabulary Building through Reading

**Writing Skill:-**Principles of Effective Writing, Structuring Essays and Reports, Writing different Audience

**Professional and Academic Communication:-**Email Etiquette and Academic and professional Settings, Writing Cover Letters and Resume, Communicating in Group projects

**B.Sc. Data Science**
**Semester-III**

**BSCDS14– Optimization Techniques**

| | | |
|---|---|---|
| **Total Marks: 100** | | **Max. Time: 3 Hrs.** |
| **End Sem Exam: 100 Marks**<br>**Min. Pass Marks: 40%** | **Internal Assessment: 50 Marks**<br>**Min. Pass Marks: 40%** | **Credits: - 2**<br>**Lectures: - 42** |

**Learning Objectives:**

The objective of this course is to teach various optimization techniques and highlight its usage indata science related 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.

**Basics of optimization: Basics** of optimization —how to formulate the problem, Maxima, minima,convex function, global solution Linear programming, simplex algorithm, Integer programming, Constraintprogramming, Knapsack problem,

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

**Introduction Genetic algorithms: Foundation** of Evolutionary theory, Evolutionary Strategies, Evolutionary, programming, Evolutionary Algorithms, Evolutionary Algorithm Case Study, Genetic Algorithm, GeneticRepresentations, Initial Population, Fitness Function, Selection and Reproduction,

**Genetic Operators:** Genetic Operators (Selection, Crossover, Mutation), Artificial Immune Systems,Other, Algorithms Harmony Search, Honey-Bee Optimization, Memetic Algorithms, Co-evolution,Multi Objective Optimization, Artificial Life, Constraint Handling

**References:**


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.Optimization Techniques, Chander Mohan and Kusum Deep, New Age Science.
3. Genetic Algorithms in Search, Optimization & Machine Learning, David E. Goldberg, Pearson EducationIndia; 1st edition(1 December 2008)
4. Genetic Algorithms: Concepts and Designs (Advanced Textbooks in Control and Signal Processing), Kim-Fung Man, Kit-Sang Tang, Sam Kwong, Springer.

**BSCDS15- Database Management System (DBMS)**

| | | |
|---|---|---|
| **Total Marks: 100** | | **Max. Time: 3 Hrs.** |
| **End Sem Exam: 100 Marks** | **Internal Assessment: 50 Marks** | **Credits: - 2** |
| **Min. Pass Marks: 40%** | **Min. Pass Marks: 40%** | **Lectures: - 42** |

**Learning Objectives:**

The objective of the course is to present an introduction to database management systems, with an emphasis onhow to organize, 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 databasedesign, relational algebra 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 ondata.
6. Familiar with basic database storage structures and access techniques: file and page organizations,indexing methods including B tree, and hashing.

**Introduction: Characteristics** of database approach, data models, database system architecture and data independence. Entity Relationship(ER) Modeling**:** Entity types, relationships, constraints.

**Relation data model:** Relational model concepts, relational constraints, relational algebra, SQL queries

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

**Transaction Processing :**ACID properties, concurrency control**,** File Structure and Indexing **,**Operations on files, File of Unordered and ordered records, overview of File organizations, Indexing structuresfor files(Primary index, secondary index, clustering index), Multilevel indexing using B and B+ trees.

**References:**


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. 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, 6thEdition, PearsonEducation,2013

## BSCDS16- Machine Learning

| Total Marks: 100 | | Max. Time: 3 Hrs. |
|---|---|---|
| End Sem Exam: 100 Marks<br>Min. Pass Marks: 40% | Internal Assessment: 50 Marks<br>Min. Pass Marks: 40% | Credits: - 2<br>Lectures: - 42 |

**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 foundationfor further understanding. This course also aims to provide 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

**Regression**:- Regression:- Simple linear Regression, Linear Regression, Multiple Regression and Polynomial Regression.

**Association Rule**: Mining Frequent Patterns, Associations, and Correlations: Basic Concepts and a Road Map, AssociationRules, the Apriori Algorithm Classification and Prediction

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

**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**

**References:**


1. Shalev-Shwartz, Shai, and Shai Ben-David. Understanding machine learning: From theory toalgorithms. Cambridge university 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.

## BSCDS17-Optimization Techniques (Python)-LAB

| Total Marks: 100 | | Max. Time: 3 Hrs. |
|---|---|---|
| End Sem Exam: 100 Marks Min. Pass Marks:40% | Internal Assessment: 50 Marks Min. Pass Marks: 40% | 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 andNatural Language Processing.

**Learning Outcomes:**

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

i. LPP Simplex method Using Pulp library Python.
ii. Knapsack implementation using python
iii. Hill climbing using Python
iv. Simulated annealing using Python
v. Genetic Implementation using Python.

**BSCDS18-DBMS – LAB**

| | | |
|---|---|---|
| **Total Marks: 100** | | **Max. Time: 3 Hrs.** |
| **End Sem Exam: 100 Marks**<br>**Min. Pass Marks: 40%** | **Internal Assessment: 50 Marks**<br>**Min. Pass Marks: 40%** | **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 andNatural Language Processing.

**Learning Outcomes:**

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

**Experiment 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

## BSCDS19-Machine Learning Lab

| | | |
|---|---|---|
| **Total Marks: 100** | | **Max. Time: 3 Hrs.** |
| **End Sem Exam: 100 Marks**<br>**Min. Pass Marks: 40%** | **Internal Assessment: 50 Marks**<br>**Min. Pass Marks: 40%** | **Credits: - 2**<br>**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 andNatural Language Processing.

**Learning Outcomes:**

1. Student will be able to understand different types of problem solving techniques.
2. Student will be able to understand and use various searching techniques.
3. 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 correctand 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**

**BSCDS21-Deep Learning**

| | | |
|---|---|---|
| **Total Marks: 100** | | **Max. Time: 3 Hrs.** |
| **End Sem Exam: 100 Marks**<br>**Min. Pass Marks: 40%** | **Internal Assessment: 50 Marks**<br>**Min. Pass Marks: 40%** | **Credits: - 3**<br>**Lectures: - 42** |

**Learning Objectives:**

The objective of this course is to provide advance knowledge of machine learning techniques. This course mainly focused onRegression 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

**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.

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

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

**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.

**References:**


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.

**BSCDS22-IoT Programming and Big Data**

| | | |
|---|---|---|
| **Total Marks: 100** | | **Max. Time: 3 Hrs.** |
| **End Sem Exam: 100 Marks**<br>**Min. Pass Marks: 40%** | **Internal Assessment: 50 Marks**<br>**Min. Pass Marks: 40%** | **Credits: - 3**<br>**Lectures: - 42** |

**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:

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.

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

**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

**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.

**Obtaining, Visualising and Analysing Data:**

Understand some methods for cleaning, summarising and visualising a large dataset; Construct and use a simplepredictive model for predicting the location of a device using signal strength and orientation. Learn how to use Python, R and R Studio to performance analysis of a large dataset;
Case studies and projects

**References:**


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

## BSCDS23- Data warehouse and Data Mining

| | | |
|---|---|---|
| **Total Marks: 100** | | **Max. Time: 3 Hrs.** |
| **End Sem Exam: 100 Marks**<br>**Min. Pass Marks: 40%** | **Internal Assessment: 50 Marks**<br>**Min. Pass Marks: 40%** | **Credits: - 3**<br>**Lectures: - 42** |

**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

**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 andProject 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 architecture3: Tool selection: Federated Data Warehouse Architecture: Apply appropriate classification and clustering techniques for data analysis

**Dimensional Modeling:** Introduction: E-R Modeling: Dimensional Modeling: E-R Modeling VS Dimensional Modeling: Data Warehouse Schemas; Star Schema, Inside Dimensional Table, Inside FactTable, 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 requirementsand 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,

**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 doesDataMining 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, BusinessIntelligence Infrastructure, Business Intelligence Applications, BI versus Data Warehouse, BI versus Data Mining, Future of BI.

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

Data Mining Techniques- An Overview : Introduction, Data Mining, Data Mining Versus DatabaseManagement 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 segmentationsoftware, 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.

**References:**


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.

## BSCDS24- Deep Learning-Lab

| Total Marks: 100 | | Max. Time: 3 Hrs. |
|---|---|---|
| End Sem Exam: 100 Marks<br>Min. Pass Marks: 40% | Internal Assessment: 50 Marks<br>Min. Pass Marks: 40% | Credits: - 3 |

**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

**BSCDS25-IoT Programming and Big Data LAB**

| | | |
|---|---|---|
| **Total Marks: 100** | | **Max. Time: 3 Hrs.** |
| **End Sem Exam: 100 Marks**<br>**Min. Pass Marks: 40%** | **Internal Assessment: 50 Marks**<br>**Min. Pass Marks: 40%** | **Credits: - 3** |

**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_Unit1_PracticalTask1.pdf

*Question 2 - Screenshots*

*Screenshot 1*


*Screenshot 2*


*Screenshot 3*

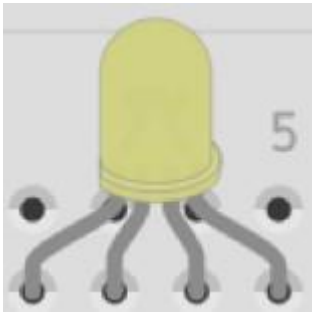
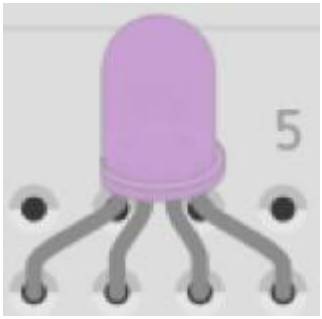
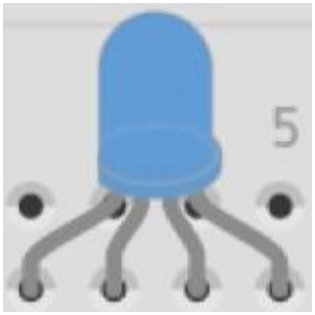
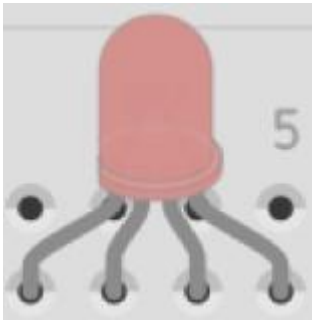

*Screenshot 4*


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 Unit 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

**BSCDS26- Data warehouse and Data Mining – LAB**

| **Total Marks: 100** | | **Max. Time: 3 Hrs.** |
|---|---|---|
| **End Sem Exam: 100 Marks**<br>**Min. Pass Marks: 40%** | **Internal Assessment: 50 Marks**<br>**Min. Pass Marks: 40%** | **Credits: - 3** |

**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 sets3 Unit-III Demonstrate performing classification on data sets

4 Unit-IV Demonstrate performing clustering on data sets5 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 BusinessAnalytics; or other data warehouse tools like Microsoft-SSIS,Informatica,Business Objects,etc.,)

A(i) 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 rowfor 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 eachchange.

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, satnavincome, 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.

**A. Sc. Data Science**

**Semester-V**

**BSCDS28- Big Data Analytics through Spark**

| | | |
|---|---|---|
| **Total Marks: 100** | | **Max. Time: 3 Hrs.** |
| **End Sem Exam: 100 Marks** | **Internal Assessment: 50 Marks** | **Credits: - 3** |
| **Min. Pass Marks: 40%** | **Min. Pass Marks: 40%** | **Lectures: - 42** |

**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 andother 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.
2. Understand the IoT network and sensor data
3. Collect and analyze large data collect through various sensor
4. Use Python for building IoT and big data based applications.

**Introduction to Spark Apache Spark Ecosystem -** Setting up the Spark Python Environment – Execution ofa PySpark Program – Resilient Distributed Datasets – Spark Architecture – Spark Project Workflow.

**Spark Programming with Python Loading and Storing Data** – Transformations – Actions – Key-ValueResilient Distributed Datasets – Local Variables – Broadcast Variables – Accumulators – Partitioning – Persistence.

**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.

**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.

**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.

**BSCDS29-Introduction to Artificial Intelligence**

| | | |
|---|---|---|
| **Total Marks: 100** | | **Max. Time: 3 Hrs.** |
| **End Sem Exam: 100 Marks** | **Internal Assessment: 50 Marks** | **Credits: - 3** |
| **Min. Pass Marks: 40%** | **Min. Pass Marks: 40%** | **Lectures: - 42** |

**Learning Objectives:**

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

**Learning Outcomes:**

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

**Introduction:** Introduction to Artificial Intelligence, Background and Applications, Turing Test and RationalAgent approaches to AI, Introduction to Intelligent Agents, their structure, behaviorand environment. Problem Solving**:** Problem Characteristics, Production Systems, Control Strategies

**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 GamePlaying, Min-Max and Alpha-Beta pruning algorithms.

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

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

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

**References:**


1. DAN.W. Patterson, Introduction to A.I and Expert Systems – PHI, 2007.
2. Russell &Norvig, Artificial Intelligence-A Modern Approach, LPE, PearsonPrentice 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

## BSCDS30-Machine Learning Operations (ML Ops)

| Total Marks: 100 | | Max. Time: 3 Hrs. |
|---|---|---|
| End Sem Exam: 100 Marks<br>Min. Pass Marks: 40% | Internal Assessment: 50 Marks<br>Min. Pass Marks: 40% | Credits: - 3 |
| | | Lectures: - 42 |

**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:**

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

**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.

**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 toCreate Deployments, Services and Scaling, Updating Deployments, Rolling Updates, Blue- Green Deployments, Canary Deployments, Managing Deployments, Jobs and CronJobs, Parallel Jobs
.

**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.

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

**Training, Tuning and Serving on AI Platform**: System Create a reproducible dataset, Implement a tunable model, Buildand push a training container, Train and tune the model, Serve and query the model.

**Kubeflow Pipelines on AI Platform:** System 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, CloudBuild Configuration, Cloud Build Triggers.

**BSCDS31: Elective – I**

## BSCDS32: BIG DATA ANALYTICS THROUGH SPARK - LAB

| | | |
|---|---|---|
| **Total Marks: 100** | | **Max. Time: 3 Hrs.** |
| **End Sem Exam: 100 Marks**<br>**Min. Pass Marks: 40%** | **Internal Assessment: 50 Marks**<br>**Min. Pass Marks: 40%** | **Credits: - 2**<br>**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

## BSCDS33: Artificial Intelligence (PROLOG / Python) -LAB

| | | |
|---|---|---|
| **Total Marks: 100** | | **Max. Time: 3 Hrs.** |
| **End Sem Exam: 100 Marks**<br>**Min. Pass Marks: 40%** | **Internal Assessment: 50 Marks**<br>**Min. Pass Marks: 40%** | **Credits: - 2**<br>**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

**Project Work**

**(In accordance with Semester subjects)**

Capstone Projects (Option to Bring Your Own Project) viz.

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**

## BSCDS35: NOSQL DATABASES

| | | |
|---|---|---|
| **Total Marks: 100** | | **Max. Time: 3 Hrs.** |
| **End Sem Exam: 100 Marks**<br>**Min. Pass Marks: 40%** | **Internal Assessment: 50 Marks**<br>**Min. Pass Marks: 40%** | **Credits: - 3**<br>**Lectures: - 42** |

- 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)

NoSQL and HDFS

Hive

HBase

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

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

## BSCDS36: CLOUD COMPUTING

| | | |
|---|---|---|
| **Total Marks: 100** | | **Max. Time: 3 Hrs.** |
| **End Sem Exam: 100 Marks** | **Internal Assessment: 50 Marks** | **Credits: - 3** |
| **Min. Pass Marks: 40%** | **Min. Pass Marks: 40%** | **Lectures: - 42** |

**Introduction Evolution of Cloud Computing** –Essential Characteristics of cloud computing – Operationalmodels 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.

**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.

**Platform Engineering Cloud Native Design and Microservices**– Containerized - Dynamically orchestrateddesign – 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.

**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.

**Cloud Security Considerations** – STRIDE Threat Model - Cloud Security Challenges – Cloud specificCryptographic 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

## BSCDS37: 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

**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 - HBaseConfiguration.

**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 networkmonitoring softwares (Wireshark, Smartsniff and Winnetcap) – Mobile equipments, Transmission methods, Issues.

**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

**Hive: Introduction** - Hive Features - Hive architecture -Hive Meta store - Hive data types -56 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.

**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 itsApplications', 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, Challengesand Future Prospects, Springer, 2014.
5. Michael Minelli, Michele Chambers Ambiga Dhiraj, "Big Data, Big Analytics : Emerging BusinessIntelligence 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 basedupon analyses of their prior purchases.
**Case Study II:** GOOGLE: Working with the U.S. Centers for Disease Control, tracks when users areinputting 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 encourageself/independent and group learning)

**A. Measurable**

1. Assignments (in writing and doing forms on the aspects of syllabus content and outside the syllabuscontent. 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 groupsas teams))
4. Study projects (by very small groups of students on selected local real-time problems pertaining tosyllabus 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 DataAnalysis, 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

**BSCDS38: Elective – II**

**BSCDS39: 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 manylocations 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 atemperature 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

## BSCDS40: 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

## BSCDS41: 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: 100 MarksMin. Pass Marks: 40%** | **Internal Assessment: 50 MarksMin. Pass Marks: 40%** | **Credits: - 3**<br>**Lectures: - 42** |

**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.

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

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

**SQOOP:** Installation, importing data, Exporting data, Running, illustrations

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

**References:**


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: 100 Marks<br>Min. Pass Marks: 40% | Internal Assessment: 50 Marks<br>Min. Pass Marks: 40% | Credits: - 3<br>Lectures: - 42 |

**Introduction: Light**, Brightness adaption and discrimination, Pixels, coordinate conventions, Imaging Geometry, Perspective Projection, Spatial Domain Filtering, sampling and quantization.
Spatial Domain Filtering**: Intensity** transformations, contrast stretching, histogram equalization, Correlation andconvolution, Smoothing filters, sharpening filters, gradient and Laplacian.

**Filtering in the Frequency domain:** Hotelling Transform, Fourier Transforms and properties, FFT (Decimation in Frequency andDecimation in Time Techniques), Convolution, Correlation, 2-D sampling, Discrete Cosine Transform, Frequency domain filtering.
Image Restoration: Basic Framework, Interactive Restoration, Image deformation and geometric transformations, image morphing, Restoration techniques, Noise characterization, Noise restoration filters, Adaptive filters, Linear, Position invariant degradations, Estimation ofDegradation functions, Restoration from projections.

**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.
Image Segmentation: Boundary detection based techniques, Point, line detection, Edge detection, Edge linking, localprocessing, regional processing, Hough transform, Thresholding, Iterative thresholding, Otsu'smethod, Moving averages, Multivariable thresholding, Region-based segmentation,Watershedalgorithm, Use of motion in segmentation

**Introduction to OpenCV and Python Units: setup** OpenCV on your computer, Core Units, The Core Functionality, imgproc Unit: Image Processing, highgui Unit : High Level GUI and Media, ml Unit : Machine Learning, video Unit: Video analysis. Python Units: Pillow, PIL, scikit-image,
Introduction to Medical Image and Processing : What is medical image, file formats, processing, application of medical image processing, case studies.

**References:**


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: 100 Marks**<br>**Min. Pass Marks: 40%** | **Internal Assessment: 50 Marks**<br>**Min. Pass Marks: 40%** | **Credits: - 3** |
| | | **Lectures: - 42** |

**Learning Objectives:**

The objective of this course is to introduce the fundamentals of Natural Language Processing and Textmining 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

**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-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 FreeGrammars - 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

.

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

**Information Extraction: IE** Application - Entity Extraction - IE Systems - Phrase Extraction. Search Engines: Early Search EnginesIndexing text for Search-Indexing Multimedia – Queries - Searching an index - Viewingsearch results.Web Mining: Web Structure - Search Engine

Coverage - A distributed Search- Crawlers Visualization Summarization:Training a summarizer - Sentence SelectionInformation Monitor.

**References:**


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 StatisticalNatural Language Processing by,MIT Press, 1999
3. Steven Bird, Ewan Klein and Edward Loper Natural Language Processingwith 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 BusinessApplications Approach", Secondimpression, 2011

## Annexure –I: Syllabus for Elective-II papers

### 1. HEALTH ANALYTICS

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

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

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

**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 inHealthcare.

**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.

**References**


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

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

## 2.Time Series Analysis and Forecasting

| Total Marks: 100 | | Max. Time: 3 Hrs. |
|---|---|---|
| End Sem Exam: 100 MarksMin. Pass Marks: 40% | Internal Assessment: 50 MarksMin. Pass Marks: 40% | Credits: - 3<br>Lectures: - 42 |

**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 ableto:

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.

**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

**Time Series Models** : White noise Process, Random walk, MA, AR, ARMA andARIMA 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 roottest, augmented Dickey – Fuller test.

**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.

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

**References:**


1. Box, G. E. P. and Jenkins, G. M. (1976). Time Series Analysis – Forecasting andControl, Holden – day, SanFrancisco.
2. Chatfield, C. (2003) Analysis of Time Series, an Introduction, CRCPress.
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

### 3.Product Development

| Total Marks: 100 | | Max. Time: 3 Hrs. |
|---|---|---|
| End Sem Exam: 100 MarksMin. Pass Marks: 40% | Internal Assessment: 50 MarksMin. Pass Marks: 40% | Credits: - 3 |
| | | Lectures: - 42 |

**Learning Objectives:**

The objective of this course is to train students on software development life cycle and familiarize them with various stage of development by giving example through various case studies.

**Learning Outcomes:**

After completion of this course successfully the students will be ableto:

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.

**Software Development Life Cycle, Building web client / Front End:**HTML5, CSS, Bootstrap, Flex, JavaScript, (React, vue.js etc.)

**Building Android client:** Android Studio setup and Basic of Android Programming, how to use various API

**Server Side Technologies (Flask and Django):**System setup of Flask and Django, learning basic of Flask and Django, Learning REST APIdevelopment, Testing API using Postman.

**Deployment and Maintenance:** Cloud-based and individual web server deployment, Continuation deployment and integration(Travis CI, GitLab CI)
**Minimum two case studies:** One web application and one Android application. Mustuse Machine learning and Data science.

**References:**


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

**B.Sc. Data Science**

**Semester-VII**

## BSCDS42: GENERATIVE AI-I

**This syllabus covers a comprehensive journey through the world of Generative AI, from understanding the basics of prompt engineering to advanced techniques, multimodal applications, coding, and practical product development. Each Unit builds upon the previous one, gradually equipping students with the skills to create their own AI-powered projects.**

**Introduction to Generative AI and ChatGPT:** Understanding Generative AI: Concepts and applications.
Evolution of Language Models: From rule-based systems to large language models.
Introduction to ChatGPT: Features, capabilities, and use cases.
Navigating Ethical Considerations: AI-generated content and responsible usage.

**Prompt Engineering and Basic Techniques:** Basics of Prompt Engineering: Crafting effective prompts for desired outputs. Generating Text with ChatGPT: Single-sentence completions and short responses.Fine-tuning Concepts: Overview of fine-tuning and transfer learning.Workshop: Creating basic prompts for specific tasks**.**

**Advanced Prompting Strategies:** Chain-of-Thought Prompting: Generating coherent multi-turn responses. Zero- and Few-shot Learning: Leveraging model's general knowledge and adapting to specific tasks. Prompt Injunctions: Controlling model behavior through instructions. Prompt Parameter Tuning: Adjusting parameters to influence output style.Workshop: Applying advanced techniques to real-world scenarios.

**Multimodal Generative AI:** Introduction to Multimodal Models: Combining text and images for richer outputs.Stable Diffusion and Mid Journey Models: Generating images and text together.
Use Cases in Creative Industries: Design, photography, and multimedia content generation. Workshop: Creating prompts for multimodal outputs.

**Applications in Code Generation and Data Science:** Code Generation with LLMs: Writing code snippets for various programming tasks. Data Science Applications: Using LLMs for data preprocessing, analysis, and modeling.Introduction to Copilot: Collaborative coding with AI assistance.
Workshop: Generating code and conducting data tasks with AI.

**Building AI-powered Applications:** Product Development Fundamentals: Design thinking, user experience, and iteration.Integrating AI in Applications: Incorporating ChatGPT or similar models. Web Application Development with Flask: Creating a GPT-powered web app.Final Projects: Students develop their own AI-enabled applications**.**

## BSCDS43 - REINFORCEMENT LEARNING

**This revised course structure includes six Units that cover both classical and deep reinforcement learning, along with a reinforcement learning project focused on enhancing cab driver recommendations. Each Unit provides a comprehensive understanding of the topics and practical application through the assignment and project.**

**Introduction to Classical Reinforcement Learning:** Understanding Reinforcement Learning: Definition, goals, and applications. Key Examples: AlphaGo and Boston Dynamics' robots. Dynamic Programming: Basics of solving RL problems using DP. Monte Carlo Methods: Introduction to episodic prediction and control. Q Learning: Learning action-value functions and optimizing policies. State and Action Value Functions: Concepts and calculations.

**Classic RL Assignment Introduction: Overview of the tic-tac-toe challenge.**

Classical RL Algorithms and Tic-Tac-Toe Agent: Recap of Dynamic Programming and Monte Carlo Methods. In-depth Q Learning: Exploring the Q Learning algorithm. Training the Agent: Implementing a Q Learning agent for tic-tac-toe. Evaluation Metrics: Defining success criteria for the agent. Fine-Tuning Strategies: Enhancing the agent's performance.
Classic RL Assignment: Developing and testing the tic-tac-toe agent.

**Introduction to Deep Reinforcement Learning: Basic Concepts of Deep Learning**: Neural networks and their applications.Merging RL and Deep Learning: Motivation and advantages. Deep Q Learning: Overview and combining Q Learning with neural networks.Policy Gradient Methods: Learning policies directly using gradient ascent. Actor-Critic Method: Combining policy-based and value-based approaches.

**Deep Reinforcement Learning Algorithms Deep Q Learning in Depth:** Working principles and training process. Policy Gradient Methods: Understanding policy gradients and their training.Actor-Critic Approach: Exploring the actor-critic architecture and training. Challenges of Deep RL: Exploring potential issues and mitigations.

**Reinforcement Learning Project Introduction**: Project Scope: Enhancing cab driver recommendations using RL. Business Context: Understanding the importance of driver retention. Cab Aggregator Service: Platform overview and recommendation system**.**

**RL Project Implementation and Deployment:** Vanilla Deep Q Learning (DQN): Overview and application for the project. Defining the Reward Structure: Designing rewards to maximize driver profits. Implementing DQN: Coding the algorithm for the recommendation problem. Model Evaluation: Testing and validating the RL-based recommendation system.
Deployment and Impact: Deploying the solution and measuring its impact on driver retention.

## BSCDS44 RESEARCH METHODOLOGY

**Introduction to Research and its Importance:** Understanding Research: Definition and scope. Importance of Research: Role in advancing knowledge and solving problems.Criticism in Research: Understanding its significance and impact. Peer Reviews: Role of peer reviews in ensuring research quality.

**Types of Research and Research Designs Descriptive vs. Analytical Research:** Distinguishing between different researchapproaches.
Applied vs. Fundamental Research: Exploring practical vs. theoretical research goals.
Quantitative vs. Qualitative Research: Understanding data collection and analysis methods. Bayesian vs. Frequentist Approach: Contrasting two statistical paradigms.

**Research Process and Literature Review Research Question**: Formulating clear and focused research questions. Hypothesis and Aims: Establishing hypotheses and research objectives. Formulating a Problem: Identifying and defining research problems. Literature Review: Evaluating existing literature and identifying gaps.

**Research Project Management and Planning:** Project Cycle: Understanding the different stages in a research project. Project Requirements on Data: Gathering and preparing necessary data. Identifying Milestones: Setting project milestones for tracking progress. Project Timelines: Using tools like Gantt Charts to plan project schedules.

**Report Writing and Presentation:** Art of Writing Papers: Strategies for effective scientific writing. Parts of a Paper: Understanding the structure of a research paper. Tools for Writing Papers: Exploring software and tools for paper writing.Publishing Papers: Submission to journals and presenting at seminars.

**Scientific Ethics and Citation:** Citation Methods and Rules: Proper citation practices and avoiding plagiarism.Honor Code: Upholding academic integrity and ethical research behavior. Research Claims: Ensuring validity and reliability of research claims. This structured course covers various aspects of research methodology, from understanding different research types to project management, report writing, and ethicalconsiderations. It equips learners with the knowledge and skills required to conduct research effectively and ethically in diverse fields.

## BSCDS45 GENERATIVE AI -I LAB

1. Craft a Python program that generates creative stories using a language modelbased on user-defined prompts.
2. Develop a script that illustrates the historical evolution of language models, creating a timeline visualization using a Python plotting library.
3. Build a chatbot application that engages in ethical content generation while interacting with users in a responsible manner.
4. Design a Python function to craft effective prompts for language models, enhancing the relevance of generated outputs.
5. Create a script that provides concise answers to user questions using ChatGPTfor a specific topic or domain.
6. Develop a notebook demonstrating fine-tuning of a language model using a small dataset, comparing its performance with the original model.
7. Construct a chat application simulating coherent multi-turn conversations with the AI, maintaining context with ChatGPT.
8. Implement a Python program that uses zero-shot learning to generate responses to various user queries on a range of topics.
9. Create an interactive tool for prompt parameter tuning, allowing users to adjust parameters like temperature and max tokens to influence output style.
10. Build a Python application that generates marketing content for a photography business, combining textual descriptions and sample images using multimodal AI models.

**BSCDS46 Reinforcement Learning- LAB**

1. Develop a Python program that applies dynamic programming to solve a grid navigation problem, showcasing the agent's optimal path.
2. Implement the Q Learning algorithm in Python to create an AI agent capable of playing Tic-Tac-Toe and learn optimal strategies.
3. Craft a command-line-based Tic-Tac-Toe game where players can compete against an AI agent trained using Q Learning.
4. Create a Python function to calculate state and action value functions in a reinforcement learning context and visualize the results.
5. Build a Python script that illustrates the evolution of language models,discussing key examples like AlphaGo and Boston Dynamics' robots.
6. Enhance the Q Learning Tic-Tac-Toe agent by implementing fine- tuning strategies and comparing its performance against baselineagents.
7. Develop a user interface where users can play against the trained Tic-Tac-Toe agent, observing its learning progress over multiple games.
8. Write a Python program that computes evaluation metrics to measurethe agent's success in learning optimal Tic-Tac-Toe strategies.
9. Create a visualization tool that showcases the Q Learning Tic-Tac-Toe agent's performance improvement over training episodes.
10. Design a Python script that introduces Monte Carlo methods andexplains their role in episodic prediction and control.

**Subject Code: BSCDS48**

**Subject Name:** Massive Open Online Courses (MOOCs)

**Teaching Scheme:** Credit 2 **ExaminationScheme:** Certificate Submission

1. Introduction to Machine Learning: https://nptel.ac.in/courses/106/106/106106139/
2. Machine Learning: https://nptel.ac.in/courses/106/106/106106202/
3. Machine Learning for Science and Engineering applications: https://nptel.ac.in/courses/106/106/106106198/
4. Introduction to Machine Learning: https://nptel.ac.in/courses/106/105/106105152/
5. Deep Learning (Part-I): https://nptel.ac.in/courses/106/106/106106184/
6. Deep Learning: https://onlinecourses.nptel.ac.in/noc19_cs54/preview
7. Naive Bayes from Scratch: https://courses.analyticsvidhya.com/courses/naive-bayes
8. Getting Started with Neural Networks: https://courses.analyticsvidhya.com/courses/getting-started- with-neural-networks
9. Machine Learning – Offered by Stanford Online - https://www.coursera.org/learn/machine-learning
10. Microsoft Exam DA-100: Analyzing Data with Microsoft Power BI
11. Microsoft Exam PL-300: Microsoft Power BI Data Analyst.

Microsoft Exam**:** Microsoft Certified: Azure Data Scientist Associate

**B.Sc. Data Science**

**Semester-VIII**

## BSCDS49 Generative AI – II

**This division maintains the flow of concepts and progresses from fundamental concepts to advanced techniques, tools, and applications, ending with a glimpse into the future of Generative AI.**

**Introduction to Embedding Large Documents with LLMS:** Explore the fundamentals of embeddings and their significance in the context of large documents. Begin the journey of building custom LLMs by understanding how to integrate databases with GenAI models.

**Storing and Indexing Embeddings of Large Documents with Vectorstores:** Dive deeper into embedding techniques and leverage vectorstores like Pinecone to store and index extensive documents and datasets. Enhance ChatGPT's contextual understanding, minimize hallucinations, and improve accuracy on data-specific tasks.

**LangChain and its Applications:** Recognize the limitations of standalone LLMs and discover the potential of LangChain. Learn how LangChain can overcome these limitations by integrating GenAI models withspecific data pools, opening up new application avenues.

**LangChain Agents, Tools, and Vectorstores for Storage and Retrieval**:Delve into the components that constitute LangChain, such as Models, Prompts, Indexes, Chains, Memory, and Agents. Gain insight into how these components collectively contribute to building robust GenAI models**.**

**Connecting Components and Leveraging Tools in LangChain:** Understand the intricacies of connecting various components within LangChain through chains. Explore the toolkit offered by LangChain, utilizing inbuilt tools to maximize the potential of your GenAI models.

**Scaling and Deploying Generative AI Apps Using Azure OpenAI Services,Future Developments:** Learn to deploy your generative AI models using Azure OpenAI services. Gain insights into the considerations involved in scaling generative AI models effectively. Conclude the course by exploring the future landscape of AI, including risk mitigation, reinforcement learning from human feedback (RLHF) as a product, and the trajectory ofmultimodal learning.

## BSCDS50 DATA ENGINEERING

**Fundamentals of Data Management and Cloud Computing:** Data management concepts,Introduction to relational database modelingBasics of cloud computing Introduction to Amazon Web Services (AWS) setup

**Big Data Fundamentals and Data Ingestion:** Introduction to Big Data conceptsHadoop framework overview MapReduce programming basics,Data ingestion techniques with Apache Sqoop and Apache Flume

**Querying and Processing with Hadoop Ecosystem**: Introduction to Hive for querying,Optimizing Spark for large-scale data processing Introduction to NoSQL databases,Exploring Apache HBase for NoSQL storage

**Cloud Infrastructure and Advanced AWS Services:** Deep dive into AWS cloud infrastructure AWS services for data processing and storageExploring AWS Lambda and AWS Glue Introduction to serverless computing

**Advanced Big Data Technologies:** In-depth study of Apache Flink,Real-time stream processing with Apache Flink Introduction to analytics using PySpark,Hands-on practice with PySpark.

**NoSQL Databases and Beyond:** Further exploration of NoSQL databasesDetailed focus on MongoDB, Advanced topics in MongoDB (e.g., sharding, indexing)Integration of NoSQL databases in data pipelines.

**BSCDS51 Data Visualization-Tableau**

This structured approach to the course allows for a gradual progression from foundational concepts to more advanced techniques, including integration with R and real-world case studies to apply Tableau analytics in practical scenarios

**Introduction to Tableau Analytics:** Tableau Introduction: Understanding Tableau's role in data visualization and analysis. Data Connection: Connecting data sources to Tableau for analysis. Calculated Fields and Hierarchy: Creating calculated fields and hierarchical structures in Tableau. Parameters, Sets, Groups: Exploring parameters, sets, and groups for dynamicvisualization control**.**

**Advanced Visualization Techniques in Tableau:** Various Visualization Techniques: Exploring different visualization types and their applications. Map-based Visualization: Creating geographical visualizations using Tableau's mapping features. Reference Lines and Totals: Adding reference lines and calculating totals and subtotals.Advanced Formatting: Utilizing advanced formatting options to enhance visual appeal.

**Data Manipulation and Analysis in Tableau: Combining Fields:** Using combined fields to merge and analyze data from differentsources. Filters and Sorting: Applying filters and sorting options to analyze specific subsets ofdata. Table Calculations: Performing calculations on the data within Tableau. Creating Dashboards: Building interactive dashboards to visualize and analyze data.

**Advanced Techniques and Integration:** Action Filters, Using action filters to create interactive connections betweenvisualizations. Creating Stories: Constructing data-driven stories by weaving together visualizations. Clustering and Time Series Analysis: Applying clustering and time series analysis techniques using Tableau.

**Integrating R and Advanced Analytics:** R Integration with Tableau: Integrating R code to enhance data analysis capabilities. Creating Statistical Models: Building statistical models with dynamic inputs using R and Tableau. Visualizing R Output: Displaying R-generated output within Tableau visualizations**.**

**Real-time Case Studies**

Case Study 1: Real-time project involving Twitter Data Analytics.
Case Study 2: Real-time project focused on Google Finance data analysis.
Case Study 3: Real-time project using IMDB Website data for analysis.
.

## BSCDS52 GENERATIVE AI-II LAB

1. Develop a Python program that simulates a grid navigation problem using dynamic programming. Implement a visualization to showcase the agent's optimal path.

2. Build a Python-based Tic-Tac-Toe game where players can compete against an AI agent trained using the Q Learning algorithm. Test the AI's performance against human players.

3. Create a Python script that calculates state and action value functions for a simple reinforcement learning scenario. Visualize these functions to understand their impact on decision-making.

4. Craft a presentation explaining the evolution of language models, using examples like AlphaGo and Boston Dynamics' robots to illustrate their impact on the AI field.

5. Enhance the Q Learning Tic-Tac-Toe agent by incorporating fine-tuning strategies. Compare the agent's performance with and without these strategies through experimentation.

6. Develop a user-friendly interface for playing Tic-Tac-Toe against a Q Learning agent. Track the agent's learning progress over a series of games.

7. Implement a Python program that calculates evaluation metrics to quantifythe success of the Q Learning Tic-Tac-Toe agent's training. Analyze the results to draw insights.

8. Create a visualization tool using Python libraries to depict the Q Learning Tic-Tac-Toe agent's performance improvement over training episodes.

9. Build a neural network in a deep learning framework to predict Q values for reinforcement learning tasks. Train and evaluate this network on a simple RL problem.

10. Implement a policy gradient method, such as REINFORCE, using a neuralnetwork for a simple RL environment. Visualize the agent's learning progress and policy changes.

**BSCDS53 -Data Engineering -LAB**

1 Write a MapReduce program that counts the frequency of words in a text file.Test the program with sample input data.

2 Use Hadoop streaming to implement a MapReduce job using Python. Calculatethe average length of words in a text file.

3 Import a specific table from a PostgreSQL database using Sqoop. Specify customdelimiters and verify the imported data.

4 Configure an Apache Flume agent to tail a log file and transfer the data to HDFS.Confirm the data is ingested and available in HDFS.

5 Create a Hive table from a Parquet file containing employee data. Write a Hivequery to find the highest-paid employee in each department.

6 Load data from a CSV file into a partitioned Hive table. Write a query thatretrieves data from a specific partition.

7 Develop a PySpark application that reads JSON data from a file, performs datacleansing, and calculates summary statistics.

8 Optimize a PySpark job by using broadcast variables and Spark's built-infunctions. Measure the performance improvement.

9 Set up an AWS Lambda function that triggers when an object is uploaded to anS3 bucket. Write a Lambda function to resize images upon upload.

10 Use AWS CloudFormation to deploy a multi-tier architecture consisting of anEC2 instance, RDS database, and an S3 bucket.

11 Implement an Apache Flink job that reads streaming data from a Kafka topic,applies windowed aggregation, and outputs results to a sink.