Maharashtra Education Society's

Abasaheb Garware College, Pune.

(Autonomous)

(Affiliated to Savitribai Phule Pune University)

Two Years M.Sc. Degree Program in Computer Science
(Faculty of Science and Technology)

Syllabi under Autonomy
M.Sc.(Data Science)

Choice Based Credit System Syllabus

To be implemented from Academic Year 2023-2024

Preamble

Master of Science in Data Science is two-year full-time program. In this program the students are encouraged to work on huge datasets, explore different case-studies related to big data, solve real world problems and analyse large volume of data. Along with this they will be taught core subjects also. This program includes studying of predictive models, statistical analysis of data, and visualization of data.

First year of the program provides foundation of Data Science. In the first year the basic subjects such as Mathematics, Statistics, Data mining Python for Data Science will be taught. Students will be trained on Python and Statistical analysis tools. This will be accompanied with advanced subjects like Machine Learning, Big Data Analysis, cloud computing. The students are provided with the elective subjects in each semester, which will allow them to select a subject of their interest. Students can do a project as an elective

Second year curriculum is designed with specialization approach which would introduce the students to Deep Learning, Data Visualization and Artificial Intelligence thereby increasing their level of expertise. An elective of Research Methodology is provided in third semester which will be useful for the students who are interested in pursuing research. Last semester is reserved for Internship or Industrial Training, where students get opportunity to work with industry experts on live project.

Throughout the course, students are encouraged to keep practical approach, work on Projects in Machine Learning and Big Data as well as publish their research work. The intention behind developing the syllabus is to make students industry ready with a researcher mind, while also focusing on overall development.

Course Objectives :

1. To enable students to develop critical thinking and problem-solving skills.
2. To provide learners with a deep knowledge of key areas of Data Science
3. To develop a researcher's mind and encourage new ideas from students
4. To enable learners conduct independent research and analysis in the field of data analytics.
5. To provide the learner with a comprehensive platform for career development, innovation and further study
6. To prepare learners for higher positions in the IT industries.

Eligibility :

Any candidate who has passed Bachelor's Degree in B.Sc. any subject, BCA, BCS, BE, B.Tech from any recognized University approved by UGC / AICTE with 50% of Marks (45% marks aggregate in case of candidate backward class categories and persons with disability)

Structure of Syllabus

Year	Semester	Course Type	Course Code	CourseTitle	Remark	Credit	No. of hours to be conducted
I	I	Core	DS-501-MJ	Exploratory Statistics and Inference		4	60
			DS-502-MJ	Mathematics for Data Science		4	60
			DS-503-MJ	Python for Data Science		4	60
			DS-504-MJP	Lab course based on all major theory courses		2	4 per batch
		Elective	DS-510-MJ	Relational Database Management System		2	30
			DS-511-MJP	Lab course based on Relational Database Management System		2	4 per batch
			OR				
			DS-512-MJ	Cloud Computing		2	30
			DS-513-MJP	Lab course based on Cloud Computing		2	4 per batch
			OR				
			DS-514-MJ	Data Mining and Data Warehousing		2	30
			DS-515-MJP	Lab course based on Data Mining and Data Warehousing		2	4 per batch
			DS-541-MN	Research Methodology		4	60
	II	Core	DS-551-MJ	Statistical Methods		4	60
			DS-552-MJ	Data Analytics		4	60
			DS-553-MJ	Machine Learning		4	60
			DS-554-MJP	Lab course based on all major theory courses		2	4 per batch
		Elective	DS-560-MJ	Design and Analysis of Algorithms		2	30
			DS-561-MJP	Lab course based on Design and Analysis of Algorithms		2	4 per batch
			OR				

			SD-562-MJ	Big Data		2	30
			DS-563-MJP	Lab course based on Big Data		2	4 per batch
			DS-581-FP	Project		4	-
II	III	Core	DS-601-MJ	Artificial Intelligence		4	60
			DS-602-MJ	Deep Learning		4	60
			DS-603-MJ	Data Visualization		4	60
			DS-604-MJP	Lab course based on all major theory courses		2	4 per batch
		Elective	DS-610-MJ	Predictive Analysis		2	30
			DS-611-MJP	Lab course based on Predictive Analysis		2	4 per batch
			OR				
			DS-612-MJ	Business Intelligence		2	30
			DS-613-MJP	Lab course based on Business Intelligence		2	4 per batch
			DS-631-RP	Research Project		4	-
	IV	Core	DS-651-MJP	Industrial Project Design		4	-
			DS-660-MJ	MOOC1		2	-
			DS-661-MJ	MOOC2		2	-
			DS-681-OJT	Industrial Training / Industrial Project		14	-

Semester I

DS-501-MJ : Exploratory Statistics and Inference

No. of Lectures : 60 (credits 4)

Unit 1 : Summary Statistics

1.1 Scale of measurement- Nominal scale, ordinal scale, Interval scale, ratio scale, Likert scale, discrete and continuous variables, Introduction of data, data types (primary, secondary, time series, cross-sectional, directional, survival, longitudinal, panel data).

1.2 Concept of Central Tendency and its measures: Mean, Median, Mode, trimmed mean, weighted and unweighted mean, combined arithmetic mean, Partition Values: quartiles, deciles, percentiles, Box Plot

1.3 Concept of Dispersion and its measures: absolute and relative measures, Range, Interquartile Range, Quartile Deviation, mean deviation about averages, mean square deviation, Variance, Standard Deviation (SD), Coefficient of variation, combined variance and SD for two or more groups. Use of relative measures for comparison of two or more datasets, suitability or choice of descriptive statistics. Five number summary and construction of boxplot.

1.4 Notion of moments, Raw and central moments, concept of symmetry and skewness, types of skewness, assessing skewness from box plot, measures of skewness, comparison of datasets with respect to type and extent of skewness.

1.5 Kurtosis: Concept of Kurtosis, Measures of Kurtosis

(All topics to be covered for raw data with large sample size and large dimension using R software. Manual calculations are not expected.)

Unit 2 : Basics of Probability

2.1 Sample space and events, Probability - classical definition, probability models, axiomatic definition of probability, probability of an event.

2.2 Concept and definition of independence of two events.

2.3 Concepts and definitions of conditional probability, multiplication theorem $P(A \cap B) = P(A).P(B|A)$, real life applications of conditional probability in data science and machine learning.

2.4 Concept of Posterior probability, problems on posterior probability.

2.5 Bayes' theorem (without proof)

2.6 Definition of sensitivity of a procedure, specificity of a procedure. Application of Bayes' theorem to design a procedure for false positive and false negative.

2.7 Numerical problems related to real life situations.

Unit 3 : Introduction to Random Variables

3.1 Notion of discrete random variable and continuous random variable.

3.2 Concept of Discrete and Continuous probability distributions (p.m.f. and p.d.f.)

3.3 Cumulative Distribution function (C.D.F.) and its properties.

3.4 Expectation, variance, moments, moment generating function (MGF), cumulant generating function (CGF), probability generating function (PGF) and its properties.

3.5 Numerical problems related to real life situations

Unit 4 : Standard Probability Distributions

4.1 Degenerate distribution, Bernoulli distribution, Binomial Distribution, Hypergeometric distribution

4.2 Discrete Uniform Distribution

4.3 Poisson Distribution

4.4 Negative Binomial Distribution and Geometric Distribution as a special case

4.5 Continuous Uniform Distribution

4.6 Exponential Distribution

4.7 Normal Distribution

4.8 Log Normal Distribution

4.9 Gamma Distribution

4.10 Weibull Distribution

4.11 Pareto Distribution

(For all the probability distributions its pmf or pdf, plotting of p.m.f. or p.d.f. curve for various parameter values, cdf, mean, variance, moments, skewness, kurtosis (without derivation) p-p plot, q-q plot, computation of probabilities and generation of random samples using R software is expected.)

Unit 5 : Correlation and Regression

5.1 Notion of Bivariate data.

5.2 Concept of Correlation, types of correlation - Positive Correlation, Negative correlation, no Correlation, covariance, Measures of correlation

- Scatter diagram
- Karl Pearson's coefficient of correlation (r),

5.3 Properties of correlation coefficient, limits of r ($-1 \leq r \leq 1$), interpretation of r, Coefficient of determination (R^2)

5.4 Meaning of regression, difference between correlation and regression. Assumptions of linear regression, Fitting of (i) simple linear regression line Y = a+bX+ where (ii) second degree curve (iii) Exponential curve (iv) Logistic curve. Introduction to multiple linear regression.

5.5 Concept of residual plot and mean residual sum of squares. Residual diagnostic-Standardized and Studentized residuals, PRESS statistic and Cook's D.

5.6 Variance inflation Factor (VIF), Identification and solution to Multicollinearity

5.7 Evaluation of the Model using R^2 and Adjusted R^2, Deciding the best fit, Variable selection methods-forward selection, backward elimination and stepwise regression.

(All topics to be covered for raw data using R software. Manual calculations are not expected.)

Unit 6 : Logistic Regression

6.1 Introduction to logistic regression, assumptions, univariate and multivariate logistic regression model. Difference between linear and logistic regression.

6.2 How to build logistic regression model in R?, null deviance, residual deviance, deciding the best model using Akaike Information Criteria (AIC) and Bayesian Information criteria (BIC).

6.3 Odds ratio in logistic regression. Real life examples of logistic regression.

(All topics to be covered for raw data using R software. Manual calculations are not expected.)

Reference Books :

1. Introduction to Linear Regression Analysis, Douglas C. Montgomery, Elizabeth A. Peck, G. Geoffrey Vining, Wiley, 2003
2. Modern Elementary Statistics, Freund J.E., Pearson Publication, 2005.
3. Fundamentals of Applied Statistics (3rd Edition), Gupta and Kapoor, S.Chand and Sons, New Delhi, 1987.
4. An Introductory Statistics, Kennedy and Gentle.
5. Statistical Methods, G.W. Snedecor, W.G. Cochran, John Wiley & sons

DS-502-MJ : Mathematical Foundations for Data Science

No. of lectures : 60 (credits 4)

Objectives :

1. Introduce concepts in linear algebra and to use it as a platform to model physical
2. problems.
3. Provide techniques for analytical and numerical solutions of linear equations and introduce the
4. concept of convergence.
5. Introduce some of the mathematical structures, concepts and notations used in discrete mathematics.
6. Introduce some concepts from graph theory, partially ordered sets, Boolean algebras.

Outcomes :

1. Students will be able to effectively use matrix algebra tools to analyse and solve systems of linear equations.
2. Students will be able to use some numerical methods to solve linear systems of equations
3. Students will be able to work with some of the mathematical structures, concepts and notations used in discrete mathematics
4. Students will be able to apply the concepts of sets, functions, relations and graph theoretic concepts to problems in computer science

Unit 1 : Matrices, System of equations, determinants and inverse of a matrix

1.1 Matrix Algebra-Row-reduced echelon form of a matrix, Inverse of a matrix

1.2 System of linear equations, Consistency and inconsistency of system of linear equations

Unit 2 :Vector spaces and Linear transformations

2.1 Vector space, subspace and span of a set, Linear dependence and independence of a set of vectors, basis and dimension

2.2 Linear transformation, rank and nullity

Unit 3 : Eigenvalues and Eigenvectors

3.1 Eigenvalues

3.2 Eigenvectors

Unit 4 : Numerical linear algebra

4.1 Gauss elimination with partial pivoting and scaling

4.2 Iterative methods for solving linear system of equations

Unit 5 : Matrix Eigenvalue Problems

5.1 Eigenvalue problems in linear system of equations

5.2 Power method for finding the dominant eigenvalue

Unit 6 : Sets, Functions and Relations, Boolean Algebra

6.1 Introduction to set theory, set relations, set operators, cardinality of sets, Cartesian product of sets

6.2 Fundamentals of functions – range, domain, injection, surjection, bijection of functions

6.3 Fundamentals of relations, reflexive, symmetric and transitive properties in relations, representing relations, applications of relations, equivalence relations, partial order relations, lattices.

6.4 Boolean functions, representing Boolean functions

Unit 7 : Graph Theory

7.1 Introduction to graph theory, directed and undirected graphs, handshaking theorem, special graph structures, graph representations and isomorphism of graphs, connectedness, components, Euler, Hamilton paths and cycles

7.2 Trees, binary trees, binary search tree, spanning trees, minimum spanning trees – Prim's and Kruskal's algorithms.

Reference Books :

1. Kenneth H. Rosen, Discrete Mathematics and its Applications, Tata McGraw Hill, 7th Ed., 2011
2. K Hoffman and R Kunze, Linear Algebra, Pearson Education, 2nd Edition, 2005

DS-503-MJ : Python for Data Science

No. of Lectures : 60 (Credits 4)

Prerequisites: Basic knowledge of any programming language.

Course Objectives:

1. Tointroduce programming concepts using python.
2. Student should be able to develop Programming logic using python.
3. To develop basic concepts and terminology of python programming.
4. To test and execute python programs.

Learning Outcomes: On completion of this cours, students will be able to:

1. Develop logic for problem solving.
2. Determine the methods to create and develop Python programs by utilizing the data structures like lists, dictionaries, tuples and sets.
3. To be familiar about the basic constructs of programming such as data, operations, conditions, loops, functions etc.
4. To write python programs and develop a small application project

Unit1:An Introduction to Python

1.1 Introduction to Python- History, features, Applications, Installing Python, Running Simple Python program

1.2 Standard data types - basic, none, Boolean (true & False), numbers, Variables, Constants

1.3 Identifiers and reserved words

1.4 Lines and indentation, multi-line statements and Comments

1.5 Input/output with print and input functions

1.6 Basic operators- Arithmetic, comparison, assignment, bitwise logical

1.7 Membership operators (in & not in)

1.8 Identity operators (is & is not)

1.9 Operator precedence

Unit2:Control Statements

2.1 Conditional Statements: if, if-else, nested if-else

2.2 Looping- for, while, nested loops, loop control statements (break, continue, pass)

2.3 Strings: declaration, manipulation, special operations, escape character, string formatting operator, Raw String, Unicode strings, Built-in String methods

Unit3: Data Structures and Functions

3.1 Lists- Concept, creating and accessing elements, updating & deleting lists, traversing a List, reverse Built-in List Operators, Concatenation, Repetition, In Operator, Built-in List functions and methods, list as stack and queue

3.2 Tuples : Tuples, Accessing values in Tuples, Tuple Assignment, Tuples as return values, Variable-length argument tuples, and Basic tuples operations, Concatenation, Repetition, in Operator, Iteration, Built-in tuple functions, indexing, slicing and matrices.

3.3 Dictionaries :Creating a Dictionary, Accessing Values in a dictionary, Updating

Dictionary, Deleting Elements from Dictionary, Properties of Dictionary keys, Operations in Dictionary, Built-In Dictionary Functions, Built-in Dictionary Methods.

3.4 Sets- Definition, transaction of set(Adding, Union, intersection), working with sets

3.5 Functions- Definitions and Uses, Function Calls, Type Conversion Functions, Math Functions, Composition, Adding New Functions, Flow of Execution, Parameters and Arguments, Variables and Parameters, Stack Diagrams, Void Functions, Anonymous functions Importing with from, Return Values, Boolean Functions, More Recursion, Functional programming tools - filter(), map(), and reduce(),recursion, lambda forms.

Unit4:Object Oriented Programming

4.1 Concept of class, object and instances

4.2 Constructor, class attributes and destructors, Inheritance, overlapping and overloading operators

4.3 Adding and retrieving dynamic attributes of classes

4.4 Programming using Oops support

Unit5:Modules, Working with files, Exception handling

5.1 Modules: Importing module, Creating & exploring modules, Math module, Random module, Time module Packages: Importing package, creating package, examples

5.2 Working with files: Creating files and Operations on files (open, close, read, write), File object attributes, file positions, Listing Files in a Directory, Testing File Types, Removing files and directories, copying and renaming files, splitting pathnames, creating and moving directories Regular Expression- Concept of regular expression, various types of regular expressions, using match function.

5.3 Exception Handling: Built-in Exceptions, Handling Exceptions, Exception with Arguments, User-defined Exceptions.

Unit6:Database

6.1 Python MySQL Database Access

6.2 Install the MySQLdb and other Packages

6.3 Create Database Connection

6.4 CREATE, INSERT, READ, UPDATE and DELETE Operation

6.5 DML and DDL Oepration with Databases

6.6 Performing Transactions

6.7 Handling Database Errors

Unit7: Web Scraping

7.1 Basics of scraping,

7.2 Scrape HTML Content From a static / dynamic pages

7.3 Parsing HTML code using packages like request and Beautiful soup

Reference Books:

1. Beginning Programming with Python for Dummies Paperback – 2015 by John Paul Mueller
2. Object-oriented Programming in Python, Michael H. Goldwasser, David Letscher, Pearson Prentice Hall-2008
3. Introduction to Problem Solving with Python by E balguruswamy,TMH publication2016

DS-504-MJP : Lab course based on Exploratory Statistics and Inference and Python for Data Science

(credits 2)

Assignments :

Exploratory Statistics and Inference

(Practical to be performed using any statistical software)

1. Diagrammatic and Graphical Representation of Statistical Data and its interpretation (considering time series, cross-sectional, directional, survival, longitudinal, panel data).
2. Summary statistics (central tendency and dispersion) (for ungrouped data).
3. Measures of Skewness and Kurtosis, Box plot.
4. Applications of conditional probability and Bayes' theorem. Examples of Bayes' theorem to design a procedure for false positive and false negative.
5. Applications of discrete distributions.
6. Applications of continuous distributions.
7. Scatter diagram, correlation coefficient (ungrouped data). Fitting of line of regression.
8. Simple regression analysis and diagnostics (Residual diagnostic- Standardized and Studentized residuals, PRESS statistic and Cook's D).
9. Multiple regression analysis and diagnostics (Residual diagnostic- Standardized and Studentized residuals, PRESS statistic and Cook's D , Variable selection methods)
10. Logistic regression (Odds ratio).

Python for Data Science

1. Assignments Based on Basics of python
2. Assignments Based on String
3. Assignments Based on Data Structures and Functions
   - List, Tuples, Sets, and Dictionary
   - Functions
4. Assignments Based on Object Oriented Programming
5. Assignments Based on Modules and File Handling
6. Assignments Based on Exception Handling
7. Assignments Based on Database Handling
8. Assignments Based on Web Scraping

DS-510-MJ : Relational Database Management Systems

No. of Lectures : 30(Credits 2)

Prerequisites:

- Basic Knowledge of Files

Course Objectives:

1. To understand the fundamental concepts of database.
2. To understand designing of database with normalization forms.
3. To understand creations, manipulation and querying of data in databases.
4. Provide an overview of the concept of NoSQL technology.

Learning Outcomes: On completion of this cours, students will be able to:

1. Design a relational database.
2. Implement SQL: Data definition, constraints, schema, queries and operations in SQL.
3. Define, compare and use the types of NoSQL Databases.

Unit1:Introduction to DBMS **2**

- 1.1 Introduction
- 1.2 Levels of abstraction & data independence
- 1.3 Structure of DBMS
- 1.4 Users of DBMS
- 1.5 Advantages of DBMS

Unit2: Conceptual Database Design Concept **4**

- 2.1 Constraints - Key constraints, Integrity constraints, Referential integrity, Unique constraint, Null/Not Null constraint, Domain constraint, Check constraint, Mapping constraints
- 2.2 Keys Concept with Examples: Primary Key, Candidate Keys and Super Keys
- 2.3 Mapping Cardinality
- 2.4 Structure of Relational Databases (Concepts of a table)

Unit3:SQL **8**

- 3.1 Introduction to query languages
- 3.2 Basic structure
- 3.3 DDL Commands
- 3.4 DML Commands
- 3.5 Forms of a basic SQL query (Expression and strings in SQL)
- 3.6 Aggregate functions
- 3.7 Nested Subqueries
- 3.8 Examples on SQL

Unit4:Relational Database Design **8**

- 4.1 Concept of Functional Dependencies
- 4.2 Armstrong's axioms

4.3 Closure of an attribute set
4.4 Algorithm to derive Primary Key and Super key with examples
4.5 Concept of Decomposition - Lossless join, Lossy join, Dependency Preservation
4.6 Concept of normalization, Normal Forms (1NF,2NF and 3NF, BCNF), Examples

Unit5:Introduction to transaction management **2**

5.1 Definition of transaction
5.2 Properties of transaction
5.3 State of transactions

Unit6:Introduction to NoSQL **6**

6.1 Brief History of NoSQL Databases
6.2 NoSQL Database Features
6.3 Difference between RDBMS and NoSQL
6.4 Why NoSQL?
6.5 Types of NoSQL Database
- 6.5.1 Key Value Databases
- 6.5.2 Document Databases
- 6.5.3 Column Family Databases
- 6.5.4 Graph Databases

6.6 When should NoSQL be Used?
6.7 NoSQL Database Misconceptions

Reference Books

1. Database System Concepts by Henry F. Korth, Abraham Silberschatz, S.Sudarshan, Tata McGraw-Hill Education publication
2. Database Management Systems by RaghuRamakrishnan, Mcgraw-hill higher Education publication
3. Beginning Databases with PostgreSQL: From Novice to Professional by Richard Stones, Neil Matthew Apress publication
4. Practical Postgresql By Joshua D. Drake, John C Worsley O'Reilly publication
5. NoSQL Distilled by Pramod Sadalge, Martin Fowler, Publication: Pearson Education
6. NoSQL for Dummies - A Willy Brand

DS-511-MJP : Lab Course on Relational Database Management Systems (Credits 2)

Assignments :

1. To create simple tables with the primary key constraint (as a table level constraint & as a field level constraint) (include all data types). Inserting data in the tables.
2. To create one or more tables with following constraints
    a. Primary Key
    b. Foreign Key
    c. Check constraint
    d. Unique constraint
    e. Not null constraint
3. To drop a table, alter schema of a table, insert / update / delete records using tables created in previous Assignments. (use simple forms of insert / update / delete statements)
4. To query the tables using simple form of select statement
    1. Select <field-list>
    from table
    [where <condition> order by <field list>]
    2. Select <field-list, aggregate functions >
    from table
    [where <condition> group by <> having <> order by <>]
5. To query tables using nested queries (use of except, exists, not exists, all clauses)

DS-512-MJ : Cloud Computing

No. of Lectures : 30 (Credits 2)

Course Objectives:

1. To learn the principles and paradigm of Cloud Computing.
2. To appreciate the role of Virtualization Technologies.
3. To develop ability to design and deploy Cloud Infrastructure.

Learning Outcomes: On completion of the course student will be able to

1. Understand core concepts of cloud computing paradigm.
2. Get into system, network and storage virtualization and outline their role in enabling the cloud computing system mode.
3. Apply fundamental concepts in cloud infrastructures to understand the trade-offs in power, efficiency and cost.

Unit 1: Introduction to Cloud Computing **8**

1.1 Overview
1.2 Layers and Types of Cloud
1.3 Desired Features of a Cloud
1.4 Benefits and Disadvantages of Cloud Computing
1.5 Cloud Infrastructure Management
1.6 Infrastructure as a Service Providers
1.7 Platform as a Service Providers
1.8 Multitenant Technology
1.9 Cloud-Enabling Technology: Broadband Networks and Internet Architecture,
1.10 Data Centre Technology, Virtualization Technology
1.11 Infrastructure as a Service, Platform as a Service, Software as a Service, Cloud
1.12 Deployment Models

Unit 2: Abstraction and Virtualization **7**

2.1 Introduction to Virtualization Technologies
2.2 Load Balancing and Virtualization
2.3 Understanding Hypervisors
2.4 Virtual Machines Provisioning and Manageability Virtual Machine Migration Services
2.5 Provisioning in the Cloud Context Virtualization of CPU, Memory, I/O Devices, Virtual Clusters and Resource management

Unit 3: Programming, Environments and Applications **8**

3.1 Features of Cloud and Grid platforms
3.2 Programming Support of Google App Engine
3.3 Programming on Amazon AWS and Microsoft Azure
3.4 Emerging Cloud Software Environments
3.5 Applications: Moving application to cloud

3.6 Microsoft Cloud Services
3.7 Google Cloud Applications
3.8 Amazon Cloud Services
3.9 CI/CD pipelines for microservices
3.10 Cloud Applications (Social Networking, E-mail, Office Services, Google Apps, Customer Relationship Management)

Unit 4: Relational Database Design **7**

4.1 Security Overview – Cloud Security
4.2 Challenges and Risks
- 4.2.1 Software-as-a-Service Security
- 4.2.2 Security Governance
- 4.2.3 Risk Management
- 4.2.4 Security Monitoring
- 4.2.5 Security Architecture Design
- 4.2.6 Data Security
- 4.2.7 Application Security
- 4.2.8 Virtual Machine Security

4.3 Identity Management and Access Control
4.4 Disaster Recovery in Clouds

Reference Books:

1. Cloud Computing: Technologies and Strategies of the Ubiquitous Data Centre by Brian J.S. Chee and Curtis Franklin Publication: CRC Press, ISBN :9781439806128
2. Mastering Cloud Computing: Foundations and Applications Programming by Rajkumar Buyya, Christian Vecchiola, S. ThamaraiSelvi, Publication: McGraw Hill, ISBN: 978 1259029950, 1259029956
3. Distributed and Cloud Computing, From Parallel Processing to the Internet of Things by Kai Hwang, Geoffrey C Fox, Jack G Dongarra, Publication: Morgan Kaufmann Publishers, 2012.

DS-513-MJP : Lab Course based on Cloud Computing

(Credits 2)

Assignments :

1. Working and Implementation of Infrastructure as a service - VM
2. Working and Implementation of Software as a service - APPSHEET
3. Working and Implementation of Platform as a service – SQL Cloud
4. Practical Implementation of Storage as a Service – GCP & AWS/AZURE
5. Working of Google drive to make spreadsheet and notes.
6. Working and Implementation of identity management - GCP
7. Write a program for web feed.
8. Execute the step to Demonstrate and implementation of cloud on single sign on - AWS
9. Create a sample mobile application using Amazon Web Service (AWS) account as a cloud service. Also provide database connectivity with implemented mobile application.
10. Installing and Developing Application Using Google App Engine – Cloud RUN
11. Installation and configuration of cloud Hadoop and demonstrate simple query.

DS-514-MJ : Data Mining and Data Warehousing

No. of Lectures : 30 (Credits 2)

Prerequisites: Basic knowledge of Applied Mathematics and Databases

Course Objectives:

1. To introduce students to the basic concepts and techniques of Data Mining and Data Warehousing
2. To study data mining algorithms for solving practical problems

Learning Outcomes: On completion of this cours, students will be able to:

1. Identify the key processes of data mining, data warehousing and knowledge discovery
2. Design data warehouse with dimensional modeling and apply OLAP operations.
3. Identify appropriate data mining algorithms to solve real world problems
4. Compare and evaluate different data mining techniques like classification, prediction, clustering and association rule mining

Unit1: Introduction to Data Mining **5**

1.1 Basic Data Mining Tasks.

1.2 DM versus Knowledge Discovery in Databases.

1.3 Data Mining Issues.

1.4 Data Mining Metrics.

1.5 Social Implications of Data Mining.

1.6 Overview of Applications of Data Mining.

Unit2: Introduction to Data Warehousing **4**

2.1 Architecture of DW

2.2 OLAP and Data Cubes

2.3 Dimensional Data Modeling-star, snowflake schemas

2.4 Data Preprocessing – Need, Data Cleaning, Data Integration &

2.5 Transformation, Data Reduction

2.6 Machine Learning, Pattern Matching

Unit3: Data Mining Tasks **5**

3.1 Frequent item-sets and Association rule mining: Apriori algorithm,

3.2 Use of sampling for frequent item-set, FP tree algorithm

3.3 Graph Mining: Frequent sub-graph mining, Tree mining, Sequence Mining.

Unit4: Classification And Prediction **8**

4.1 Decision tree learning

- 4.1.1 Construction, performance, attribute selection
- 4.1.2 Issues: Over-fitting, tree pruning methods, missing values,
- 4.1.3 Continuous classes
- 4.1.4 Classification and Regression Trees (CART)

4.2 Bayesian Classification:

4.2.1 Bayes Theorem, Naïve Bayes classifier,
4.2.2 Bayesian Networks
4.2.3 Inference
4.2.4 Parameter and structure learning
4.2.4.1 Linear classifiers.
4.2.4.2 Least squares, logistic, perceptron and SVM classifiers
4.3 Prediction
4.3.1 Linear regression
4.3.2 Non-linear regression

Unit5: Accuracy Measures. **2**

5.1 Precision
5.2 Recall
5.3 F-measure
5.4 confusion matrix
5.5 cross-validation
5.6 bootstrap

Unit6: Clustering **3**

6.1 k-means
6.2 Expectation Maximization (EM) algorithm
6.3 Hierarchical clustering, Correlation clustering

Unit7: Software for data mining and applications of data mining **3**

7.1 R and Weka tool.

Reference Books:

1. Data Mining: Concepts and Techniques, Han, Elsevier ISBN:9789380931913/9788131205358
2. Margaret H. Dunham, S. Sridhar, Data Mining – Introductory and Advanced Topics, Pearson Education
3. Tom Mitchell, —Machine Learning||, McGraw-Hill, 1997
4. R.O. Duda, P.E. Hart, D.G. Stork. Pattern Classification. Second edition. John Wiley and Sons, 2000.
5. Christopher M. Bishop, —Pattern Recognition and Machine Learning||, Springer 2006
6. Raghu Ramkrishnan, Johannes Gehrke, Database Management Sysstems, Second Edition, McGraw Hill International
7. Ian H.Witten, Eibe Frank Data Mining: Practical Machine Learning Tools and Techniques, Elsevier/(Morgan Kauffman), ISBN:9789380501864

DS-515-MJP : Lab Course Based on Data Mining

(Credits 2)

Assignments:

1. Assignments Based on Using the WEKA Workbench
2. Assignments based Basic classification and usage of weka
3. Assignments based on Classification
4. Assignments based on regression

SEMESTER II

DS-551-MJ : Statistical Methods

No. of Lectures : 60 (credits 4)

Unit 1 : Sampling

1.1 Population, Introduction to Sampling, probability sampling and nonprobability sampling

1.2 probability sampling

- 1.2.1 Simple random Sampling (SRSWR and SRSWOR)
- 1.2.2 Systematic random sampling
- 1.2.3 Stratified Random Sampling
- 1.2.4 Cluster Sampling
- 1.2.5 Two stage sampling

1.3 Nonprobability sampling

- 1.3.1 Convenience sampling
- 1.3.2 Consecutive sampling
- 1.3.3 Purposive sampling
- 1.3.4 Quota sampling
- 1.3.5 Snowball sampling

1.4 Concept of Sampling Error, non-sampling error, Formation of strata and clusters using special criterion and hence drawing a sample by stratified sampling and cluster sampling method.

Unit 2 : Sampling Distributions

2.1 Introduction to Sampling distributions, testing normality of number of samples by Shapiro Wilk test.

2.2 Chi square distribution

2.3 Student's t distribution

2.4 Snedecor's F distribution

2.5 Interrelations among t, chi-square and F distributions

2.6 Central Limit Theorem (Various Versions) and its applications.
Monte Carlo simulations from the sampling distributions using standard normal distribution.

Unit 3 : Testing of hypothesis

3.1 Notion of parameter, parameter space, introduction to problem of estimation, Point estimation: definition of an estimator and estimate, illustrative examples. Mean Square Error (M.S.E.) of an estimator, definitions and illustrations of statistic, mapping of statistic to parameter (sample to population), sampling distribution of a statistic, standard error of a statistic.

3.2 Concept of hypothesis, statistical hypothesis, types of hypotheses: simple and composite, null and alternative hypotheses with notations used. Critical region, type - I and type - II errors, level of significance, p – value, one sided and two-sided tests.

3.3 Large Sample Tests -Tests for population means using critical region approach, p - value approach and two sided confidence interval approach.

3.4 Tests based on t distribution

Tests for population means:

3.4.1 Single sample with unknown variance and two sample for unknown equal variances (tests for one-sided and two-sided alternatives)

3.4.2 two sided confidence interval for population mean and difference of means of two independent normal populations

3.4.3 Paired t-test for one-sided and two-sided alternatives

3.4.4 test for significance of population correlation coefficient

3.5 Tests based on Chi-square distribution:

3.5.1 For independence of two attributes: 2 x 2 and r x s contingency table

3.5.2 For goodness of fit

3.6 Tests based on F-distribution:

3.6.1 Test for significance of two independent sample variance by verifying all the assumptions of F-test

(All topics to be covered for raw data using R software. Manual calculations are not expected.)

Unit 4 : Analysis of Variance

4.1 Concept of ANOVA

4.2 Classification of ANOVA- One Way and Two Way, verifying the assumptions of one-way ANOVA

4.3 Application of ANOVA to test the overall significance of regression model.

4.4 Bartlett's test for homogeneity of variance (homoscedasticity)

4.5 Kruskal Wallis test for non-normal sample

(All topics to be covered using R software. Manual calculations are not expected.)

Unit 5 : Time Series

5.1 Meaning and Utility of time series.

5.2 Components of Time Series, Exploratory data analysis: Time series plot to (i) check any trend and seasonality in the time series (ii) identify the nature of trend

5.3 Additive and Multiplicative models.

5.4 Methods of estimating trend: moving average method, least squares, method of exponential smoothing, Holt Winters method (single, double and triple), Choosing parameters for smoothing and forecasting based on exponential smoothing

5.5 Elimination of trend using additive and multiplicative models.

5.6 Simple time series models: AR(P) model, p=1,2,… ,Case studies of real life Time Series: Price index series, share price index series, economic time series: temperature and rainfall time series, wind speed time series, pollution levels.

5.7 Introduction to ARIMA Modelling.

5.8 Case study – analysis of real life time series data (report of case study to be submitted)

Reference Books :

1. Fundamentals of Applied Statistics (3rd Edition), Gupta and Kapoor, S.Chand and Sons, New Delhi,
2. Time Series Analysis,4th Edition, Box and Jenkin, Wiley.
3. Modern Elementary Statistics, Freund J.E., Pearson Publication,
4. Common Statistical Tests, Kulkarni M.B., Ghatpande S.B., Gore S.D., Satyajeet Prakashan, Pune,
5. Time Series Methods, Brockell and Devis, Springer

DS-552-MJ : Data Analytics

No of lectures 60 (Credits 4)

Prerequisites :

- Basic of mathematics and statistics
- Basic programming Knowledge of python
- Knowledge of databases Course

Objectives :

1. Deploy the Data Analytics Lifecycle to address data analytics projects.
2. Develop in depth understanding of the key technologies in data analytics.
3. Apply appropriate analytic techniques and tools to analyse data, create models, and identify insights that can lead to actionable results.

Course Outcomes : On completion of the course, student will be able to

1. Use appropriate models of analysis, assess the quality of input, and derive insight from results.
2. Analyse data, choose relevant models and algorithms for respective applications
3. Understand different data mining techniques like classification, prediction, clustering and association rule mining
4. Apply modelling and data analysis techniques to the solution of real world business problems

Unit 1 : Introduction to Data Analytics

1.1 Concept of data analytics

1.2 Data analysis vs Data analytics

1.3 Types of analytics Diagnostic Analytics, Predictive Analytics , Prescriptive Analytics, Exploratory Analysis, Mechanistic Analysis

1.4 Mathematical models - Concept Model evaluation: metrics for evaluating classifiers - Class imbalance - AUC, ROC (Receiver-Operator Characteristic) curves, Evaluating value prediction models

Unit 2 : Machine Learning Overview

2.1 Introduction to Machine Learning, deep learning

2.2 Artificial intelligence

2.3 Applications for machine learning in data science

2.4 The modelling process - Engineering features and selecting a model, Training the model, Validating the model, Predicting new observations

2.5 Types of machine learning - Supervised learning, Unsupervised learning, Semi-supervised learning, ensemble techniques

2.6 Regression models - Concept of classification, clustering and reinforcement learning.

Unit 3 : Mining Frequent Patterns, Associations, and Correlations

3.1 Types of patterns can be mined

3.2 Class/Concept Description: Characterization and Discrimination, Mining Frequent Patterns, Associations, and Correlations, Classification and Regression for Predictive Analysis, Cluster Analysis, Outlier Analysis
3.3 Mining frequent patterns - Market Basket Analysis.
3.4 Frequent Itemsets, Closed Itemsets, and Association Rules
3.5 Frequent Itemset Mining Methods
3.6 Apriori Algorithm
3.7 Generating Association Rules from Frequent Itemsets
3.8 Improving efficiency of apriori algorithm
3.9 Frequent pattern growth (FP-growth) algorithm

Unit 4 : Social Media Analytics

4.1 Overview of social media analytics - Social Media Analytics Process, Seven layers of social media analytics, accessing social media data
4.2 Important social media analytics methods
4.3 Social network analysis - Link prediction, Community detection, Influence maximization, Expert finding, Prediction of trust and distrust among individuals
4.4 Challenges to social media analytics

Unit 5 : Text Analytics

5.1 Introduction to Natural Language Processing
5.2 Text Analytics : Tokenization, Bag of words, Word weighting : TF-IDF, n-Grams, stop words, Stemming and lemmatization, synonyms and parts of speech tagging Sentiment Analysis
5.3 Document or text summarization
5.4 Trend analytics

References :

1. Data Science Fundamentals and Practical Approaches, Gypsy Nandi, Rupam Sharma, BPB Publications, 2020.
2. The Data Science Handbook, Field Cady, John Wiley & Sons, Inc, 2017
3. Data Mining Concepts and Techniques, Jiawei Han, Micheline Kamber, Jian Pei, Morgan Kaufmann, Third Edition, 2012.
4. A Hands-On Introduction to Data Science, Chirag Shah, University of Washington Cambridge University Press
5. The Data Science Design Manual, Steven S. Skiena, Springer, 2017
6. Introducing data science: big data, machine learning, and more, using Python tools

DS-553-MJ : Machine Learning

No. of Lectures : 60(Credits 4)

Prerequisites:

- Familiarity with Probability Theory, Multivariable Calculus, Linear Algebra
- Programming in Python (NumPy, SciPy, Pandas, Matplotlib, Seaborn, SciKit-Learn,StatsModel)

Course Objectives:

1. To introduce students to the basic concepts and techniques of Machine Learning.
2. To write python programs using machine learning algorithms for solving practical problems.
3. To understand about Machine Learning Library and use cases.
4. To understand about the process of deploying ML model.

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

1. Recognize the characteristics of machine learning that make it useful to real-world problems.
2. Process available data using python libraries and predict outcomes using Machine Learning algorithms to solve given problem.
3. Able to estimate Machine Learning models efficiency using suitable metrics.
4. Design application using machine learning techniques.

Unit1:Introduction to Machine Learning

1.1 Introduction: Machine Learning, Examples of Machine Learning applications, Training versus Testing, Positive and Negative Class, Cross-validation.
1.2 Types of Learning: Supervised, Unsupervised and Semi-Supervised Learning.
1.3 Components of Generalization Error (Bias, Variance, underfitting, overfitting)
1.4 Metrics for evaluation viz. accuracy, scalability, squared error, precision and recall, likelihood, posterior probability
1.5 Dimensionality Reduction: Introduction to Dimensionality Reduction, Subset Selection, Introduction to Principal Component Analysis

Unit2: Supervised Machine Learning

2.1 Regression- Linear Regression, Univariate Regression, Multivariate Linear Regression, Polynomial Regression, Logistic Regression
2.2 Classification - K – Nearest Neighbours (KNN), Naive Bayes Theorem, Support Vector Machine, Decision Tree, Random Tree
2.3 Model Accuracy, Confusion Matrix

Unit3: Un-Supervised Machine Learning

3.1 Clustering Fundamentals
3.2 K-means
3.3 Hierarchical Clustering (Agglomerative, Divisive),
3.4 Dendrogram
3.5 Selecting optimal number of clusters: Within Clusters Sum of Squares (WCSS) by Elbow Method
3.6 Association Rules - Support, Confidence and Lift
3.7 Apriori Algorithm

Unit 4. Reinforcement Learning

4.1 Introduction- Environment, State, Reward, Policy, Value
4.2 Upper Confidence Bound
4.3 Thompson Sampling
4.4 Q-Learning

Unit 5. Deep Learning

5.1 Introduction to Deep Learning
5.2 Artificial Neural Network
5.3 Convolution Neural Network
5.4 Recurring Neural Network
5.5 Generative Adversarial Networks

Reference Books:

1. Mitchell, Tom M. "Machine learning. WCB." (1997).
2. Rogers, Simon, and Mark Girolami. A first course in machine learning. CRC Press, 2015.
3. Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. The elements of statistical learning. Vol.1. Springer, Berlin: Springer series in statistics, 2001.
4. Witten, Ian H., and Eibe Frank. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2005.
5. Machine learning course material by Andrew Ng, Stanford university
6. Sutton, Richard S., and Andrew G. Barto. Reinforcement learning: An introduction. Vol.
   No. 1. Cambridge: MIT press, 1998.
7. Iba, Takashi, et al. "Learning patterns: A pattern language for active learners." Conference on Pattern Languages of Programs (PLoP). 2009.
8. Nikhil Buduma, “Fundamentals of Deep Learning”, O‟REILLY publication, second edition 2017, ISBN: 1491925612
9. Josh Patterson, Adam Gibson, “Deep Learning: A Practitioners Approach”, O‟REILLY,
   SPD, ISBN: 978-93-5213-604-9, 2017 Edition 1st.

DS-554-MJP : Lab course based on Statistical Methods, Data Analytics and Machine Learning

(Credits 2)

Assignments :

Statistical Methods

Practical to be performed using any statistical software

1. Testing normality of number of samples by Shapiro Wilk test.
2. Central Limit Theorem (Various Versions) and its applications.
3. Monte Carlo simulations from the sampling distributions (Chi square , Student's t and Snedecor's F) using standard normal distribution.
4. Large Sample Tests -Tests for population means using critical region approach, p-value approach and two sided confidence interval approach.
5. Tests based on t distribution - Tests for population means:
   a. Single sample with unknown variance and two sample for unknown equal variances (tests for one-sided and two-sided alternatives)
   b. two sided confidence interval for population mean and difference of means of two independent normal populations
   c. Paired t-test for one-sided and two-sided alternatives
   d. test for significance of population correlation coefficient
6. Tests based on Chi-square distribution: For independence of two attributes: 2 x 2 and r x s contingency table.
7. Tests based on F-distribution: Test for significance of two independent sample variance by verifying all the assumptions of F-test.
8. Classification of ANOVA- One Way and Two Way, verifying the assumptions of one-way ANOVA , Application of ANOVA to test the overall significance of regression model and Bartlett's test for homogeneity of variance.
9. Methods of estimating trend: moving average method, least squares, method of exponential smoothing, Holt Winters method (single, double and triple).
10. ARIMA Modelling.

Data Analytics

1. Frequent itemset and association rule mining Load Transactional data set.
2. Linear and Logistic regression
3. Text Analytics
4. Sentiment analysis

Machine Learning Assignments:

1. Write a python program to Prepare Scatter Plot (Use Forge Dataset / Iris Dataset)
2. Write a python program to find all null values in a given data set and remove them.
3. Write a python program the Categorical values in numeric format for a given dataset.

4. Write a python program to transform data with Principal Component Analysis (PCA)
5. Write a python program to implement simple Linear Regression for predicting house price.
6. Write a python program to implement multiple Linear Regression for a given dataset.
7. Write a python program to implement Polynomial Regression for given dataset.
8. Write a python program to Implement Naïve Bayes.
9. Write a python program to Implement Decision Tree whether or not to play tennis.
10. Write a python program to implement linear SVM.
11. Write a python program to find Decision boundary by using a neural network with 10 hidden units on two moons dataset
12. Write a python program to implement k-nearest Neighbors ML algorithm to build prediction model.
13. Write a python program to implement k-means algorithm on a synthetic dataset.
14. Write a python program to implement Agglomerative clustering on a synthetic dataset.
15. Write a python program to find Decision boundary by using a neural network with 10 hidden units on two moons dataset.

DS-560-MJ : Design and Analysis of Algorithms
No. of Lectures : 30 (Credits 2)

Prerequisites: Basic knowledge of algorithms, programming concepts and Data Structures.

Course Objectives:

1. To learn basic Algorithm Analysis techniques and understand the use of asymptotic notation.
2. To select the appropriate algorithm by doing necessary analysis of algorithms.
3. Understand different algorithm design strategies.
4. Understand the use of data structures in improving algorithm performance.

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

1. Compare functions using asymptotic analysis and describe the relative performance.
2. Use the design techniques introduced i.e., dynamic programming, greedy algorithm. etc. to design algorithms for more complex problems and analyze their performance.
3. Be familiar with the major graph algorithms and their analyses.

Unit1: Basics of Algorithms **5**

1.1 Introduction Algorithm definition and characteristics

1.2 Use of asymptotic notation, Big-Oh Notation, Little-Oh, Big Omega, Little-Omega, and Theta Notations

1.3 Solving recurrence relations using Master and Substitution method

1.4 Linear Search

1.5 Bubble Sort

Unit2:Divide and Conquer Strategy **6**

2.1 Control abstraction

2.2 Searching Problem-Binary search with time complexity

2.3 Sorting Problem-Merge sort with best, average and worst-case time complexity analysis

2.4 Integer Multiplication Problem (Booth's Algorithm)

Unit3:Graph **6**

3.1 Concept and terminologies

3.2 Graph Representation –Adjacency matrix, Adjacency list, Inverse Adjacency list

3.3 Types of edges and Articulation Point

3.4 Graph Traversals – Breadth First Search and Depth First Search (with implementation)

Unit4:Greedy Method **8**

4.1 Control abstraction

4.2 Fractional Knapsack problem

4.3 Job sequencing with deadlines
4.4 Minimum-cost spanning trees: Kruskal's and Prim's algorithm
4.5 Huffman coding
4.6 Single Source Shortest Path: Dijkstra's Algorithm

Unit5:Dynamic Programming **5**

5.1 0/1 Knapsack Problem (function method and merge-and-purge)
5.2 All pairs Shortest Path (Floyd Warshall Algorithm)
5.3 Bellman ford algorithm
5.4 Travelling Salesperson problem

Reference Books:

1. Computer algorithms by Ellis Horowitz, Sartaj Sahani, Sanguthevar Rajasekaran, Publication: Galgotia Publication
2. Algorithms by T. Cormen, C. Leiserson, & R. Rivest, Publication: MIT Press
3. The Design and Analysis of Computer Algorithms by A. Aho, J. Hopcroft & J. Ullman, Publication: Addison Wesley

DS-561-MJP : Lab course based on Design and Analysis of Algorithms (Credits 2)

Assignments:

1. Assignment based on Time Complexity Computation
2. Problem solving assignment on Divide and Conquer
3. Problem solving assignment on Graph
4. Problem solving assignment on Greedy Method
5. Problem solving assignment on Dynamic Programming

DS-562-MJ : Big Data

No. of Lectures : 30 (Credits 2)

Prerequisites: Basic knowledge of Programming Language (preferably Java), Basic knowledge of SQL, exposure to Linux Environment

Course Objectives:

1. To understand Big Data platform and its Use cases
2. To provide overview of Apache Hadoop
3. To provide understanding of HDFS concepts
4. To Understand concepts of Map Reduce
5. To provide hands on Hadoop Eco System

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

1. Identify Big Data
2. List components of Hadoop and Hadoop Eco system
3. Get knowledge of Map Reduce framework
4. To use framework like Pig and Hive to process Big Data

Unit1: Introduction to Big Data **5**

1.1 What is Big Data?
1.2 History of Data Management – Evolution of Big Data
1.3 Structuring Big Data
1.4 Elements of Big Data
1.5 Big Data Analytics
1.6 Careers in Big Data
1.7 Future of Big Data
1.8 Use of Big Data in Social Networking
1.9 Use of Big Data in Preventing Fraudulent Activities
1.10 Use of Big Data in Detecting Fraudulent Activities in Insurance Sector
1.11 Use of Big Data in Retail Industry

Unit2: Introduction to Hadoop **8**

2.1 Introducing Hadoop, RDBMS versus Hadoop
2.2 Distributed Computing Challenges, History and overview of Hadoop,
2.3 Use Case of Hadoop
2.4 Processing Data with Hadoop
2.5 Interacting with Hadoop Ecosystem
2.6 HDFS (Hadoop Distributed File System)
2.7 Processing Data with Hadoop
2.8 Managing Resources and Applications with Hadoop YARN (Yet Another Resource Negotiator)

Unit3: Hadoop Distributed File System (HDFS) **4**

3.1 The Design of HDFS,

3.2 HDFS Concepts, Basic Filesystem Operations,
3.3 Hadoop Filesystems.
3.4 The Java Interface- Reading Data from a Hadoop URL
3.5 Reading Data Using the Filesystem API, Writing Data.
3.6 Data Flow- Anatomy of a File Read, Anatomy of a File Write
3.7 Limitations.

Unit4: Understanding Map Reduce Fundamentals **5**

4.1 The MapReduce Framework
4.2 Mapper, Reducer, Combiner, Partitioner
4.3 Searching, Sorting, Compression
4.4 Techniques to Optimize MapReduce Jobs
4.5 Uses of MapReduce
4.6 Role of HBase in Big Data Processing
4.7 Developing Simple MapReduce Application
4.8 Points to Consider while Designing MapReduce

Unit5: Hive **4**

5.1 Introducing Hive
5.2 Getting Started with Hive
5.3 Hive Services
5.4 Data Types in Hive
5.5 Built-In Functions in Hive
5.6 Hive DDL
5.7 Data Manipulation in Hive
5.8 Data Retrieval Queries
5.9 Using JOINS in Hive

Unit6: Analyzing Data with Pig **4**

6.1 Introducing Pig
6.2 Running Pig
6.3 Getting Started with Pig Latin
6.4 Working with Operators in Pig
6.5 Debugging Pig
6.6 Working with Functions in Pig
6.7 Error Handling in Pig

Reference Books:

1. Seema Acharya, Subhashini Chellappan, —"Big Data and Analytics", Wiley Publications, 2nd Edition,2014
2. Tom White, —"Hadoop: The Definitive Guide", O'Reilly, 3rd Edition, 2012.
3. DT Editorial Services - "Big Data, Black Book: Covers Hadoop 2, MapReduce, Hive, YARN, Pig, R and Data Visualization"

DS-563-MJP : Lab Course Based on Big Data

(Credits 2)

Assignments:

1. Assignments Based on simple Hadoop commands
2. Assignments based on Map Reduce
3. Assignments using Apache HIVE
4. Assignments based on PIG

DS-581-FP : Project
(4 credits)

Guidelines :

1. Students should work in a team of minimum 2 and maximum 3 students.
2. Students can choose a project topic without any restriction on technology or domain.
3. Students are expected to carry out the following tasks during project work –
    a. Problem Identification
    b. Literature Review/ Study
    c. Feasibility Study
    d. Design (includes DB design, system flow or design diagrams)
    e. Modelling (if applicable)
4. Track sheet will be maintained by project guide for each group separately.
5. Project guide will conduct presentation for the work done (mentioned in point no. 3)
6. Project groups will work on actual development and/or implementation of proposed idea/topic.
7. Record of progress will also be maintained by keeping track sheet.
8. At the end of the project, the group should prepare a report which should conform to international academic standards. The report should follow the style in academic journals and books, with clear elements such as: abstract, background, aim, design and implementation, testing, conclusion and full references, Tables and figures should be numbered and referenced to in the report.
9. Minimum 2 demos will be conducted for the project work.
10. The final project presentation with demonstration (EE) will be evaluated.

Evaluation Pattern

The internal and external evaluation will be 50-50%

For all the courses, which are of four credits, total marks will be 100. Out of 50 marks will be allotted for internal evaluation and 50 marks for external evaluation.

For all the courses, which are of two credits, total marks will be 50. Out of 25 marks will be allotted for internal evaluation and 25 marks for external evaluation.

Theory Courses of four credits :

- Internal evaluation will be of 50 marks for which 3 continuous evaluation exams of 15, 15 and 20 marks will be conducted
- External evaluation will be of 50 marks

Theory Courses of two credits :

- Internal evaluation will be of 25 marks for which 2 continuous evaluation exams of 15 and 10 marks will be conducted
- External evaluation will be of 25 marks

Practical Courses of four credits :

- Internal evaluation will be of 50 marks out of which 30 marks will be for assignment submissions done throughout the semester and a test/viva will be conducted for 20 marks
- External evaluation will be of 50 marks

Practical Courses of two credits :

- Internal evaluation will be of 25 marks out of which 15 marks will be for assignment submissions done throughout the semester and a test/viva will be conducted for 10 marks
- External evaluation will be of 25 marks

Methods of assessment for internal evaluation:

Seminar, objective test, open book test, Quiz, viva, projects, assignments, group discussion, research paper review, case study, industrial visit

Passing percentage

The student must secure at least 40% marks of that course to earn the full credit.

Examination	Credits	Marks Out of	Passing marks (40%)
Internal	4	50	20
External	4	50	20
Internal	2	25	10
External	2	25	10

Note: There is separate passing for internal and external examinations.