St.Xavier's College (Autonomous), Ahmedabad
Department of Data Science
M.Sc. Big Data Analytics (2025-2027)

Pre-requisites for Course Work

1. **Microsoft Excel for Data Analysis**
    a. Excel Tables, Filters, Sorting
    b. Pivot Tables and Charts
    c. Formats, Formulas, Dates
    d. Functions – Mathematical, Statistical, Text, Date

Reference:

On-line courses/Tutorials:

i. Microsoft Virtual Academy:
    a. Analyzing and Visualizing Data with Excel https://mva.microsoft.com/en-US/training-courses/analyzing-and-visualizing-data-with-excel-11157
    b. Data Analysis with Excel https://mva.microsoft.com/en-US/training-courses/data-analysis-with-excel-16654

ii. Edx.Org:
    a. Introduction to Data Analysis using Excelhttps://www.edx.org/course/introduction-to-data-analysis-using-excel-0

iii. Coursera.org:
    a. Introduction to Data Analysis Using Excel https://www.coursera.org/learn/excel-data-analysis

2. **Basic Unix Programming**
    a. Basic Unix Commands
    b. Handling files and folders
    c. Concatenation, find and replace, modify file & texts
    d. Basic summary commands

Reference:

On-line courses/Tutorials:

i. Data Camp:
    a. Introduction to Shell for Data Science https://www.datacamp.com/courses/introduction-to-shell-for-data-science

ii. Linux.Org:
    a. Linux Beginner Tutorials https://www.linux.org/forums/linux-beginner-tutorials.123/

b. Github - Organizing with Unix: https://rafalab.github.io/dsbook/organizing-with-unix.html

Book:

i. Data Science at the Command Line, Jeroen Janssens, https://www.datascienceatthecommandline.com/

3. Mathematics, Statistics and Basic Programming

1. **Basic algebra and co-ordinate geometry:** Set theory**:** Definition of set, Representation, Types, Basic operations, Venn Diagram, Applications, Cartesian coordinates, Distance between two points, Equation of straight line (slope intercept form)
2. **Functions and Graphs:** Definition of function, Types of functions, Graphical representation of functions.
3. **Elementary calculus:** Limit of a function, Properties of limits (only statements), Continuity of a function, Differentiation: Introduction, Differential coefficients of a few important functions (only results), Important results on differentiation (without proof), Maxima and Minima, Integration, Integration by substitution, Integration by parts, Definite integral, Properties of Definite Integral (without proof).
4. **Matrix Algebra:** Definition and Notation, Operations on Matrices, Determinant, Transpose, Adjoint, Inverse.
5. **Counting Basics**: Fundamental rule of counting, Factorial, Permutation and Combination
6. **Introduction to Computer and Programming**
   Introduction, Basic block diagram and functions of various components of computer. Binary Numbers, Concepts of Hardware and software. Types of software. Compiler and Interpreter. Concepts of Machine level (low level), Assembly level and High-level Programming. Flowcharts. Algorithms
7. **Basic building blocks of programming language**
   Data types, constants, variables, operators, expressions, evaluation of expressions, type conversion, precedence and associativity. Simple statements, Decision making statements, Looping statements, Nesting of control structures, break and continue, goto statement.

Big Data Analytics - COURSE STRUCTURE
wef June 2025

St Xavier's College (Autonomous), Ahmedabad
Department of Big Data Analytics
M.Sc. Big Data Analytics (2025-2026)

Program Specific Outcome : The **Master of Science (MSc.) in Big Data Analytics** programme is designed to equip graduates with advanced knowledge and practical expertise in **Big Data Analytics, Data Science and Artificial Intelligence**. The programme emphasizes analytical thinking, research, innovation, and the ethical use of data to address complex real-world challenges. Upon successful completion of the programme, graduates will be able to:

PSO1: Advanced Data Analytics and Statistical Modelling: Apply advanced statistical methods, machine learning algorithms, and data mining techniques to analyse structured, semi-structured, and unstructured datasets for extracting meaningful insights and supporting data-driven decision-making.

PSO2: Big Data Technologies and Scalable Computing

Design, develop, and deploy scalable big data solutions using distributed computing frameworks, and data engineering tools for efficient storage, processing, and analysis of large-scale datasets.

PSO3: Data Science and Artificial Intelligence Applications

Develop intelligent data-driven applications by integrating data science methodologies with artificial intelligence, deep learning, and predictive analytics to solve complex problems across diverse domains.

PSO4: Visualization and Communication

Ensure data quality, and create effective visualizations and interactive dashboards to communicate analytical findings clearly to technical and non-technical stakeholders.

PSO5: Ethical, Secure, and Responsible Data Practices

Apply ethical principles, data governance policies, privacy regulations, and cybersecurity best practices in the collection, management, analysis, and dissemination of data to ensure responsible use of information.

PSO6: Research, Innovation, and Professional Competence

Conduct independent research, innovate with emerging data science technologies, and develop industry-ready solutions through interdisciplinary collaboration, critical thinking, and lifelong learning to address real-world challenges.

SEMESTER – I

CORE Paper: STATISTICAL METHODS
Course Code: PBD-1801
No. of Credits: 04
Learning Hours: 60 hrs
Practical's to be conducted using R

Course Outcomes

- CO1 :Understand data pre-processing and data cleaning
- CO2 : Identify the suitable descriptive measures to explore the data.
- CO3 :Learn the analysis of attributes and Chi-square tests for categorical data
- CO4: Understand timeseries data and its components and learn to do exploratory analysis
- CO5: Apply basic statistical methods in real data using R

Unit No.	Name of Unit	No. of lectures
Unit-1	**Data Collection & Visualization**	
	Concepts of measurement, scales of measurement, design of data collection formats with illustration, data quality and issues with date collection systems with examples from business, cleaning and treatment of missing data	15 (9T+6P)
Unit-2	**Basic Statistics**	
	Frequency table, histogram, measures of location, measures of spread, skewness, kurtosis, percentiles, box plot, correlation and simple linear regression.	20 (12T+8P)
Unit-3	**Contingency Tables**	
	Two-way contingency tables, measures of association, testing for dependence	10 (6T+4P)
Unit-4	**Introduction to Time Series**	
	Components of time series, decomposition of time series data, Smoothing auto correlation, stationarity	15(9T+6P)

SUGGESTED BOOKS:

1. Statistics: David Freedman, Pobert Pisani & Roger Purves, WW.Norten & Co. 4th Edition 2007.
2. The visual display of Quantitative Information: Edward Tufte, Graphics Press, 2001.
3. Best Practices in Data Cleaning: Jason W. Osborne, Sage Publications 2012
4. Time Series Analysis and Its Applications: Robert H. Shumway and David S. Stoffer, Springer 2010

CORE Paper: PROBABILITY & STOCHASTIC PROCESS
Course Code: PBD-1802
No. of Credits: 04
Learning Hours: 0 hrs
Practical's to be conducted using R

- CO1 :Apply basic ideas of probability and probability distributions in real life situation
- CO 2 : Apply the concept of stochastic process in different sectors like brand switching in Marketing Analytics .
- CO 3: Understand basic models in time series data
- CO 4: Apply time series data analysis through R

Unit No.	Name of Unit	No. of lectures
Unit-1	**Basic Probability** Concepts of experiments, Outcomes, Sample space, Events, Combinatorial probability, Birthday paradox, Principle of inclusion & exclusion, Conditional probability, Independence, Bayes Theorem	15 (12T+3P)
Unit-2	**Univariate Probability Distributions** Random Variables: discrete and continuous probability models, some probability distributions: Binomial, Poisson, Geometric, Uniform, Normal, exponential	20 (14T+6P)
Unit-3	**Bivariate Probability Distributions** Bivariate Random variables, Joint probability distribution (pmf and pdf) and Cumulative Distribution Functions, Marginal probability distribution (marginal pmf and marginal pdf) and Conditional distribution Functions, Independence of random variables, Conditional mean and conditional variance.	15 (12T+3P)
Unit-4	**Stochastic Process** Markov Chains, Classification of states, Stationery distribution, limit theorems, Poisson process, illustrations and applications.	10(6T+4P)

SUGGESTED BOOKS:

1. A First Course in Probability: Shelden M. Ross, 2014.

2. Introduction to Stochastics Process: Paul G. Hoel, Sydney C. Port & Charles J. Stone, Waveland Press, 1987.

3. Time Series Analysis and Its Applications: Robert H. Shumway and David S. Stoffer, Springer 2010.

CORE Paper: LINEAR ALGEBRA & LINEAR PROGRAMMING
Course Code: PBD-1803
No. of Credits: 05
Learning Hours: 60 hrs
Practical's to be conducted using R/Python

Course Outcomes

- CO 1: Student will be able to perform matrix operations and employ fundamental concepts of matrix theory.
- CO 2: Students will be able to employ linear algebra to solve some scientific problems.
- CO 3: Student will be able to use fundamental concepts like system of simultaneous linear equations, eigenvalues and eigenvectors in some applicable concepts.
  CO 4: Student will be able to formulate and model linear programming problems.
- CO 5: Student will be able to solve real life problems using linear programming problems and interpret solution of linear programming problems.

Unit No.	Name of Unit	No. of lectures
Unit-1	**Introduction to Matrices**	
	Linear equations and matrices, matrix operations, solving system of linear equations, Gauss-Jordan method, Concept & Computation of determinant and inverse of matrix	15 (10T+5P)
Unit-2	**Linear Algebra**	
	Eigen values and Eigen vectors, Illustrations of the methods, Positive semi definite and position definite matrices, illustrations	15(10T+5P)
Unit-3	**Introduction to Linear Programming**	
	Definition of the problem, convex sets, corner points, feasibility, basic feasible solutions	15 (12T+3P)
Unit-4	**Solving Linear Programming Problems**	
	Graphical Method, Simplex Method, Artificial Variables, Big-M method	15(12T+3P)

SUGGESTED BOOKS

1. Linear Algebra and Its Application: Gilbert Strang, 4th Edition, Academic Press.
2. Hands-On Matrix Algebra Using R (Active and Motivated Learning with Applications), Hrishikesh D Vinod, World Scientific.
3. Linear Programming: G. Hadley, Addison-Wesley.

CORE Paper: COMPUTING FOR DATA SCIENCE WITH JAVA

Course Code: PBD-1804
No. of Credits: 05
Learning Hours: 60 hrs
Practical's to be conducted using Java

Course Description: This course provides a comprehensive introduction to computing concepts and tools using the Java programming language, specifically tailored for data science applications. Topics include Java programming fundamentals, data structures, algorithms, database connectivity, and software development practices.

Course Objectives:

- CO 1: To equip students with foundational computing skills using Java for data science tasks.
- CO 2: To introduce students to data structures and algorithms relevant to data manipulation and analysis.
- CO 3: To familiarize students with database connectivity and SQL queries in Java.
- CO 4: To teach students software development best practices in Java.

Unit No.	Name of Unit	No. of lectures
Unit-1	**Introduction to Java Programming** Introduction to Java programming language, Basic syntax, data types, and control structures	15 (6T+9P)
Unit-2	**Object-oriented Programming Concepts** Classes, Objects, Inheritance, Polymorphism	15(6T+9P)
Unit-3	**Database Connectivity and SQL with Java** Introduction to JDBC (Java Database Connectivity), Connecting to databases, Executing SQL queries from Java, Handling result sets, Data manipulation and retrieval using SQL	15 (6T+9P)
Unit-4	**Advanced Java Concepts** Collections, Multithreading and concurrency, Exception handling, Input/output streams	(6T+9P)

SUGGESTED BOOK:

1. Data Structures and Algorithm using Java, 6th Ed. Michael T. Goodrich and Roberto Tamassia, John Wiley & Sons, Inc

CORE Paper : DATABASE MANAGEMENT SYSTEM
Course Code: PBD-1805
No. of Credits: 04
Learning Hours: 60 hrs

- CO 1: Learn basic data models and Hadoop Ecosystem
- CO 2: Understand few relational and non-relational databases
- CO 3: Explore hands on experience on Oracle/MySql
- CO 4: Implementation of ORACLE SQL/MS SQL/MySQL.

Unit No.	Name of Unit	No. of lectures
Unit-1	**Basic Concepts** a. Introduction to Databases and Data at Scale, Data, Database, DBMS, Drawbacks of file system, Data Independence, Database system architecture, Mappings, DBMS system and its functionalities, Database Administrator. b. Database models: hierarchical Model, Network Model, Entity relationship model: entity sets, attributes, relationships, ER diagrams, mapping rule, Schema, Enhanced entity model: specialization and generalization, association, Relational Database Model: schemas, tuples, domains, keys, relationships, Operations on relational databases: Basic operations, Relational algebra.	15 (9T+6P)
Unit-2	**Relational and Non-Relational Databases** Relational database: a. Normalization and Data integrity, Functional dependencies, Keys, Normal Forms b. Query Processing: Query costs, Query optimization, Transformation of relational expression / equivalence rule, Evaluation Plans Transaction processing: ACID properties, CAP theorm, Serialization. NON-SQL DATABASES: a. Relational and object relational databases (overview) i. Need for normalization ii. Join operations as the glue between table iii. Distributed relational databases b. Key-value stores c. Document model	15(9T+6P)

d. Columnar databases
e. Graph databases

Structure, various operations, normalization, SQL, No-SQL, Graph Database, Parallel and distributed database, Map-Reduce.
Lab using SQL/Oracle/MySql for Relational databases;
Hadoop(any), MongoDB, GraphDB for Big Data

Unit-3 **Introduction to Big data: characteristics, applications**

Hadoop Architecture: Introduction, Map-Reduce Paradigm, HDFS, and Hadoop Ecosystem, Parallel and distributed data bases, File Handling: Storage, Concept of database security. 15 (9T+6P)

Unit-4 **Introduction to ORACLE**

Overview of ORACLE and SQL, Introduction, datatypes, literals, creating tables, inserting data in tables, retrieving and fetching data from table, saving work. 15(9T+6P)

CONSTRAINTS: primary key, foreign key, Unique, Check, not null, default value.

Modifying table structure: ALTER commands, TRUNCATE, DROP, PURGE, FLASH BACk, RENAME.

Operations: DML (insert, update, delete, select, row, group by, join)

Multiple table data manipulation:

i. Joins
ii. sub queries
iii. set operators
iv. row num
v. top-n.

Views: data dictionary views, User creation and management-revoke, grant.

SUGGESTED BOOKS

1. Database system concepts: Abraham Silberschartz, Henry F. Korth and S. Surarshan, McGraw Hill, 2011.
2. Hadoop 2 Quick-Start Guide: Learn the Essentials of Big Data Computing in the Apache Hadoop 2 Ecosystem, Douglas Eadline, Addison-Wesley, Pearson Education India; First edition (1 March 2016)
3. Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, EMC Education Services, 2015.

CORE Paper : PYTHON PROGRAMMING
Course Code: PBD-1806
No. of Credits: 04
Learning Hours: 60 hrs

- CO 1: Write python functions
- CO 2: Understand packages and importing packages
- CO 3: Learn file handling
- CO 4: Develop OO Programming Concepts and get exposure on Exception Handling along with OO programming
-

Unit No.	Name of Unit	No. of lectures
Unit-1	**Basic Concepts and writing Functions** Introduction to Python interpreter, Control statements, Data Types, Defining a function, calling a function, passing by value or reference, anonymous function	15
Unit-2	**File Handling and Packages** Opening and Closing Files, Reading and Writing Files, Directories in Python, What are Packages? Import package	15
Unit-3	**Exception Handling** Python errors and Built-in exceptions, user defined exceptions, exception handling	15
Unit-4	**OO Programming Concepts** OOP, class, Inheritance, overloading	15

SUGGESTED BOOKS

1. Core Python Programming: Dr. R. Nageswara Rao, DreamTech, Second Edition
2. Python for Everybody: Exploring Data in Python 3: Charles Severance
3. Python Cookbook: Recipes for Mastering Python 3: David Beazley, 3rd Edition

SEMESTER – II

CORE Paper: FOUNDATIONS OF DATA SCIENCE (PROGRAMMING FOR BIG DATA)
Course code: PBD-2801
No. of credits: 05
Learning hours: 75hrs
Practical's to be conducted using Python/ R

- CO 1 : Learn basic concept of graph theory and understand algorithm on connectedness,shortest path algorithm, spanning tree of graph and random graph which used in industries.
- CO 2: Learn High dimensional space and understand geometry of large data set.
- CO 3 :Learn above Singular value decomposition, which has application in image processing, principal component analysis etc.
- CO 4: Random walk is use to make prediction, students will learn regarding the same.
- CO 5: Learn few algorithm for massive data problems.

Prerequisites:

1. Linear Algebra
2. Basic probability and probability distribution functions.
3. Multivariable Calculus.

Unit No.	Name of Unit	No. of Lectures
Unit-1	Graph Theory Basic Concepts, Algorithms for connectedness, shortest path, Minimum Sampling Tree, Random Graph, G(n,p) model, Giant Component, Connectivity, Cycles, Non-Uniformmodels, Applications.	15
Unit-2	High Dimensional Space Properties, Law of large numbers, Sphere and cube in high dimension, Generating points on the surface of a sphere, Gaussians in High dimension, Random projection, Applications.	15
Unit-3	Singular Value Decomposition Best rank k approximation, Power method for computing the SVD, Applications.	15
Unit-4	Random Walks and Markov Chains Introduction to Random Walk and Markov Chains, Stationary Distribution, Fundamental theorem of Markov chains, Markov Chain Monte Carlo method, Metropolis-Hasting algorithm, Gibbs sampling	15

Unit-5	Algorithm for Massive Data Problems Frequency Moments of data streams, matrix algorithms.	15

SUGGESTED BOOKS:

1. Foundations of Data Science: John Hopcroft & Ravindran Kannan.
2. Network Science: Albert Laszlo Barabasi.
3. Graph theory: Reinhard Diestel
4. Graph Algorithm: Shimon Even, Guy Even
5. Graph Algorithm for Data Science by Tomaz Bratanic (Manning Publications Co).

CORE Paper : ADVANCED STATISTICAL METHODS
Course Code: PBD-2802
No. of Credits: 05
Learning Hours: 75 hrs
Practical's to be conducted using R

- CO 1: Identify best estimators by applying knowledge on properties of estimators
- CO 2: Analyze and apply statistical inference by showing how hypothesis testing can be developed for situations involving single population and two populations
- CO 3: Apply concepts of the linear models in real life situation
- CO 4: Understand basics of Bayesian statistics
- CO 5: Learn applications of Bayesian estimation in real life scenarios.

Unit No.	Name of Unit	No. of Lectures
Unit-1	**Estimation** Parameter Space, Sampling distribution of sample mean, Characteristics of estimators, Unbiasedness, Consistency, Efficiency and sufficiency, MVUE, Cramer Rao Lower Bound, MVBE, UMVUE, Maximum likelihood estimates; EM algorithm	15
Unit-2	**Test of Hypotheses** Framing of Statistical Hypothesis, One tail and two tail tests, Two types of errors, test statistic, parametric tests for equality of means & variances, Calculation of Probability of type II errors, Power of a test.	15
Unit-3	**Linear Model** Linear Models, Gauss Markov Model, least square estimators, BLUE, Introduction to Design of Experiments, Principles of Experimental design, Analysis of variance: One-way, Two-way	15
Unit-4	**Bayesian Statistics with Applications** Introduction, Bayes Rule (Revision), Conditional probability distribution, Conditional mean and variance, Difference with frequentist approach to statistical inference, Basic components of Bayesian Models (Prior and Posterior distribution, proportionality formula), Bayesian approach to point estimation (single parameter), Computational Tool: Sampling Importance Resampling. Advantages of Bayesian Models.	15
Lab	Practicals based on all 4 units	15

SUGGESTED BOOKS:

1. J. Bickel and K. A. Docksum, Statistical Inference, 2nd Edition, Prentice Hall.
2. Peter M. Lee (2012), Bayesian Statistics: An Introduction, 4th Edition, Wiley
3. George E. P. Box and George C. Tiao (1992), Bayesian Inference in Statistical Analysis, Wiley.

CORE Paper : INTRODUCTION TO MACHINE LEARNING
Course Code: PBD-2803
No. of Credits: 04
Learning Hours: 60 hrs
Practical's to be conducted using Python

- CO 1: Study the basic concepts and techniques of Machine Learning
- CO 2: Learn supervised and unsupervised algorithms
- CO 3: Evaluate the model performance
- CO 4: Learn importance of Ensemble methods
- CO 5: Understand Association rules in Machine Learning
- CO 6: Clustering algorithms and checking goodness of fit of clusters through Silhouette Analysis

Unit No.	Name of Unit	No. of Lectures
Unit-1	**Foundations of Machine Learning** Basic definitions: supervised, unsupervised, and reinforcement learning, Hypothesis space and inductive bias, Model evaluation, train-test splits, cross-validation, Overfitting and underfitting, bias-variance trade-off.,Linear regression: Normal Equations, Ordinary Least Square, Batch, Stochastic and mini-batch gradient descent, probabilistic interpretation, Decision Trees: ID3, CART, pruning.	15
Unit-2	**Classification and Dimensionality Reduction** Instance-based learning: k-NN, various distance metrics, Feature reduction methods: **PCA**: Eigen decomposition, SVD, applications in face recognition, **LDA**: Linear discriminant analysis, Fisher's criterion, Introduction to SVD and feature selection, Collaborative filtering and recommendation systems., Logistic Regression: decision boundary, maximum likelihood estimation, Naïve Bayes: Gaussian, multinomial, Laplace smoothing, handling imbalanced data	15
Unit-3	**Advanced Models** Support Vector Machines: Hard margin, soft margin, slack variables, Dual problem, kernel trick (Polynomial, RBF, Gaussian), SMO algorithm, multiclass classification, Neural Networks: Perceptron, activation functions, Multi-layer perceptron, XOR problem, Backpropagation, training strategies (batch, mini-batch, stochastic GD), Introduction to deep neural networks.	15
Unit-4	**Learning Theory and Ensemble Learning** Computational Learning Theory, PAC Learning model, sample complexity, VC Dimension, Ensemble Learning: Bagging, Boosting, Random Forests.	15

SUGGESTED BOOKS:

1. Machine Learning: Tom Mitchell
2. Pattern Recognition and Machine Learning: Christopher Bishop, Springer,2006
3. An Introduction to Statistical Learning: Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, Springer,2015.
4. Python Machine Learning: Sebastian Raschka,2015

CORE Paper : ENABLING TECHNOLOGIES FOR DATA SCIENCE I
Course Code: PBD-2804
No. of Credits: 04
Learning Hours: 60 hrs

Course Outcomes

- CO 1: Learn the concept of various big data platforms like Hadoop ecosystem and its major components.
- CO 2 :Learn NoSQL database
- CO3: Learn workflow scheduler tool Oozie in Hadoop environment.

Unit No.	Name of Unit	No. of Lectures
Unit-1	**Big data and Hadoop**: Hadoop architecture, Single node & Multi-node Hadoop, Hadoop commands, Hadoop daemon, Task instance, Hadoop ecosystem and its installation, Illustrations.	15
Unit-2	**Map-Reduce**: Framework, Developing Map-Reduce program, Life cycle method, Serialization, Running Map-Reduce in local and pseudo-distributed mode, Illustrations	15
Unit-3	**SQOOP & PIG:** Importing data, exporting data, Running, Illustrations, Schema, Commands, Illustrations.	15
Unit-4	**NoSQL database & Oozie** Features, Types, NoSQL vs. SQL, Advantages and Disadvantages, What is Oozie? Workflow, packaging and deploying an Oozie workflow application, Features.	15

SUGGESTED BOOKS

1. Hadoop The Definitive Guide : Tom White , 4th Edition, 2017

2. Hadoop in Action : Chuck Lam, 2010

3. Data-intensive Text Processing with Map Reduce : Jimmy Lin and Chris Dyer, Morgan & Claypool Publishers, 2010

CORE Paper : VALUE THINKING
Course Code: PBD-2805
No. of Credits: 02
Learning Hours: 30 hrs

Course Outcomes

- CO 1: Enhance logical thinking, argumentative logic, evidence gathering, and drawing inference from evidences.
- CO 2: Get more awareness of the factors like deep rooted prejudices, pre-conceived ideas, psychological and sociological influences that sub-consciously come into play in decision making and forming impressions.

This course involves watching few movies (list provided below) and reading few books (list provided below) that deals mostly with argumentative logic, evidence, drawing inference from evidences. After watching the movies and reading the books, there will be general discussion amongst the students. Couple of case studies that involve mostly logical thinking will also be presented. Each student will prepare a term paper. Evaluation will be on the basis of this term paper and participation in group discussion.

Movies:(1 credit)

1. Twelve Angry Men
2. Roshoman by Kurosawa
3. Trial of Nuremberg
4. Mahabharata by Peter Brook
5. Jurassic Park
6. Interstellar

Books: (1 credit)

1. The Hound of the Baskervilles by Arthur Conan Doyle
2. Five Little Pigs by Agatha Christie
3. The Purloined Letter by Edger Allan Poe
4. The Case of the Substitute Face
5. Caves of Steel by Issac Assimov
6. Fahrenheit 451 by Ray Bradbury

Case Studies:

Elective course : PRESCRIPTIVE ANALYTICS
Course Code: PBD-2906
No. of Credits: 02 (1T+2P)
Learning Hours: 45 hrs
Practical's to be conducted using R/LINDO/SOLVER

Course outcomes

- CO 1: To develop analytical and modelling skills in optimization and simulation.
- CO 2: To understand LP duality theory and its applications in sensitivity analysis and economic interpretation.
- CO 3: To introduce large-scale LP model structures and basic solution concepts.
- CO 4: To equip students with simulation tools and techniques for decision-making under uncertainty.
- CO 5: To integrate optimization and simulation for real-world prescriptive analytics problems.

Unit No.	Name of Unit	No. of Lectures
Unit-1	**Duality and Large-Scale Linear Programming Models**	
	Role of prescriptive analytics in decision science, Optimization vs. predictive analytics, Types of models: deterministic, stochastic, constrained optimization, Primal and dual formulation, Canonical forms and dual construction, Weak and strong duality, Complementary slackness conditions, Economic interpretation of dual variables, Shadow prices and sensitivity analysis, Dual simplex method (conceptual overview), Sensitivity Analysis: Change in objective function coefficients, **Large-Scale Linear Programming Basics:** Characteristics of large LPs, Decomposition principles (introductory), Column generation (basic concept)	7
	Lab works on LP modelling and solving, Sensitivity interpretation with real datasets.	15
Unit-2	**Simulation with Applications**	
	Introduction to Simulation: Simulation in prescriptive analytics, Deterministic vs stochastic simulation, Monte Carlo and Discrete Event Simulation (DES) basics, Steps in simulation modelling, when to use simulation vs optimization	8
	Probability Foundations for Simulation Random number generation, Random variate generation (normal, exponential, Poisson, uniform, empirical distributions), An Example of a Discrete-Event Simulation, Monte Carlo Simulation, An Example of Monte Carlo Simulation, Simulations with Continuous Random Variables, An Example of a Stochastic Simulation	
	Lab works on Simulation	15

TEXTBOOKS & REFERENCES

Optimization / Duality / LP

1. Hillier & Lieberman — *Introduction to Operations Research*
2. Bazaraa, Jarvis & Sherali — *Linear Programming and Network Flows*
3. Taha — *Operations Research: An Introduction*
4. Bertsimas & Tsitsiklis — *Introduction to Linear Optimization*

Simulation

1. Averill Law — *Simulation Modeling and Analysis*
2. Banks et al. — *Discrete-Event System Simulation*
3. Rossetti — *Simulation Modeling in R*
4. Matloff — *The Art of R Programming* (for simulation coding support)

CORE LAB Paper : ML LAB
Course Code: PBD-2806L
No. of Credits: 02
Learning Hours: 30 hrs
Practical's to be conducted using Python

Unit No.	Name of Unit	No. of Lectures
Unit-1	**Foundations of ML & Data Preprocessing** Handling missing values (imputation, deletion, advanced methods), Outlier detection and treatment (IQR, z-score, isolation forest),Feature scaling methods: Min-Max,Standardization, Robust scaling, Encoding categorical variables: One-hot, label encoding, target encoding. Train-test split, k-fold cross-validation, and stratified sampling,Bias-variance analysis with synthetic datasets, Linear Regression (Normal Equation & OLS) from scratch.,Gradient descent (Batch, Stochastic, Mini-batch) for regression. (from scratch and using in-built functions.)	8
Unit-2	**Classification & Dimensionality Reduction** Logistic Regression using sklearn and from scratch, k-NN classification and regression (with different distance metrics), Decision Tree Classifier & Regressor with pruning, Naïve Bayes (Gaussian & Multinomial) on text and tabular data, PCA with eigen decomposition & SVD – visualization in 2D, LDA for dimensionality reduction & Fisher's criterion, Face recognition using PCA (Olivetti dataset), Recommender system with collaborative filtering (MovieLens dataset).	7
Unit-3	**Advanced Models** Support Vector Machines: Hard-margin vs. Soft-margin. (Code from scratch, code using sk-learn and tuning hyperparameters), SVM with kernel trick (RBF, Polynomial, Gaussian) on synthetic data, SMO algorithm (implement simple version), Perceptron learning rule on linearly separable data, XOR problem: failure of perceptron, solution with MLP, Backpropagation implementation for MLP, Training strategies comparison: Batch vs. Mini-batch vs. SGD.	8
Unit-4	**Learning Theory & Ensemble Models** PAC Learning simulation: sample complexity vs. accuracy, VC Dimension estimation for linear classifiers, Bagging with Decision Trees, Boosting (AdaBoost, Gradient Boosting), Random Forest implementation & feature importance, Model comparison report: Linear, Logistic, SVM, NN, Ensemble on one dataset.	7

****

SEMESTER – III

CORE Paper : ENABLING TECHNOLOGIES FOR DATA SCIENCE-II
Course Code: PBD-3801
No. of Credits: 04
Learning Hours: 60 hrs

- CO 1: Get awareness about big data platforms (Spark, Scala, Mahout)
- CO2 : Explore clustering algorithms in Mahout
- CO3: Train, evaluate and deploy classification models

Course Overview & Course Objectives
Unit 1) Spark
Unit 2) Scala
Unit 3) Mahout

Unit No.	Name of Unit	No. of Lectures
Unit-1	**Spark** Features, Architecture, Components of Spark, Resilient Distributed Datasets – data structure of Spark, working with Key/Value pairs, Loading and Saving data, Core Programming - RDD Transformations, Actions. Executing a Spark Application.	15
Unit-2	**Scala** Features, Basic Syntax, Data types, Variables, Classes and objects, Access modifiers, operators, if construct, loop statement, functions, OOP concepts, Array, String, Exceptions, Collections, File Handling, Multithreading.	15
Unit-3	**Mahout** Features, Installation, **Recommendations**: Introducing recommenders, representing recommender data, making recommendations.	15
Unit-4	**Practical Applications** **Clustering**: exploring distance measure, data representation, clustering algorithms in Mahout, evaluating and improving clustering quality **Classification**: training a classifier, evaluating and tuning a classifier, deploying a classifier	15

SUGGESTED BOOKS:

1. Learning Spark: Holden Karau, Andy Konwinski, Patrick Wendell, and Matei Zaharia, OReilly

2. Programming in Scala: Martin Odersky, Lex Spoon, Bill Venners,2008

3. Mahout in Action: Sean Owen, Robin Anil, Ted Dunning, Ellen Friedman,2012

CORE Paper : ADVANCED MACHINE LEARNING
Course Code: PBD-3802
No. of Credits: 04
Learning Hours: 60 hrs
Practical's to be conducted in Python

Course Outcomes

- CO 1: Explain the fundamental concepts of deep learning, including neural network architectures, regularization methods, optimization techniques, and transfer learning.
- CO2: Apply Natural Language Processing (NLP) techniques for text representation, similarity analysis, and understanding modern language models.
- CO3: Analyze and implement basic computer vision and image processing techniques for feature extraction, image analysis, and object recognition tasks.
- CO 4: Describe the principles of reinforcement learning and formulate decision-making problems using agent-environment interactions and value-based learning methods.
- CO 5: Evaluate the applicability of deep learning, NLP, computer vision, and reinforcement learning techniques in solving real-world artificial intelligence problems.

Course Overview and Course Objectives

Unit 1) Deep Learning
Unit 2) Natural Language Processing
Unit 3) Computer Vision and Image Processing
Unit 4) Reinforcement Learning

Unit No.	Name of Unit	No. of Lectures
Unit-1	Deep Learning	
	Recap of MLP, Regularization techniques, Convolutional Neural Networks (basics), Optimization algorithms (SGD, Adam), Transfer learning (introductory)	15
Unit-2	Natural Language Processing	
	Text preprocessing techniques, Bag of Words and TF-IDF, Vector space model and cosine similarity, Word embeddings (Word2Vec concept), , Recurrent Neural Networks (overview), LSTM (conceptual), Attention mechanism, Transformer models (introductory)	15
Unit-3	Computer Vision and Image Processing	
	Digital image representation, Color models and image histograms, Convolution and filtering, Edge detection techniques, Image segmentation basics, Feature extraction (HOG, SIFT conceptual), Object detection concepts, CNN applications in vision.	15
Unit-4	Reinforcement Learning	
	Agent–environment framework, Reward and return, Markov Decision Process (conceptual), Policy and value functions, Bellman equation (conceptual form), Q-learning algorithm, Exploration vs exploitation, Introduction to Deep Reinforcement Learning.	15

SUGGESTED BOOKS:

1. Pattern Recognition and Machine Learning: Christopher Bishop, Springer,2006.
2. An Introduction to Statistical Learning: Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, Springer,2015.
3. Python Machine Learning: Sebastian Raschka,2015

CORE Paper : VISUAL ANALYTICS
Course Code: PBD-3803
No. of Credits: 02 (1T & 2P)
Learning Hours: 45 hrs

- CO1: **Introduce the principles of data visualization** and the role of visual perception in effectively communicating data-driven insights
- CO2: **Develop an understanding of visualization fundamentals**, data modeling techniques, and methods for comparing and contrasting data through appropriate visual representations.
- CO3: **Develop practical proficiency in Tableau** for data preparation, visualization creation, dashboard development, and interactive data exploration.
- CO4: **Equip learners with the ability to transform raw data into meaningful visual stories** that facilitate analysis, interpretation, and business decision-making.

a) *Introduction to Data Visualization and Visual Perception*

b) *Fundamentals of Visualization, Data Modeling and Compare and Contrast*

c) *Data Visualization Best Practice and Not-So-Best Practices*

d) *The Use of Color in Data Visualization and Dashboard Design*

e) *Typography and Data Visualization Design*

f) *Exam and Infographics*

g) *Interactive Data Visualization*

h) *Mapping Data*

i) Hands-on practice on Tableau.

(30P)

SUGGESTED BOOKS:

1. Information Dashboard Design: Displaying Data for At-a-glance Monitoring by Stephen Few
2. Learning Tableau 10, Joshua N. Milligan, Packt Publishing
3. Practical Tableau : 100 Tips, Tutorials, and Strategies from a Tableau Zen Master, Ryan Sleeper, O'Reilly
4. Tableau Cookbook – Recipes for Data Visualization, Shweta Sankhe-Savale
5. Tableau for Dummies, Molly Monsey, Paul Sochan
6. Tableau Dashboard Cookbook, Jen Stirrup, Packt Publishing

CORE Paper : ADVANCED MACHINE LEARNING (Lab)

Course Code: PBD-3804L

No. of Credits: 02 (4P)

Learning Hours: 30 hrs

Unit No.	Name of Unit	No. of Lectures
Unit-1	**Deep Learning** 1.Implementing and Visualizing a Multi-Layer Perceptron (MLP) from Scratch 2. Effect of Regularization Techniques: L1, L2, and Dropout in Neural Networks 3. Building and Training a Basic Convolutional Neural Network (CNN) 4. Comparative Study of Optimization Algorithms: SGD vs Adam 5. Hyperparameter Tuning and Learning Curve Analysis 6. Transfer Learning with Pretrained CNN Models (Feature Extraction Approach)	8
Unit-2	**Natural Language Processing** 1.Text Preprocessing Pipeline: Tokenization, Stopwords, and Lemmatization 2. Implementing Bag of Words and TF-IDF Vectorization 3. Vector Space Model and Cosine Similarity for Text Classification 4. Word Embeddings using Word2Vec (Pretrained Model Exploration) 5. Sequence Modeling with RNN and LSTM (Text Generation or Sentiment Analysis) 6. Introduction to Attention Mechanism and Transformer-based Text Classification	7
Unit-3	**Computer Vision and Image Processing** 1. Digital Image Representation and Histogram Analysis 2. Image Filtering and Convolution Operations (Blurring, Sharpening) 3. Edge Detection Techniques (Sobel, Canny) Implementation 4. Image Segmentation using Thresholding and Clustering 5. Feature Extraction using HOG and SIFT (Conceptual + Implementation) 6. Object Detection and CNN-based Image Classification Application	8
Unit-4	**Reinforcement Learning** 1. Simulating the Agent–Environment Interaction Framework 2. Understanding Rewards, Returns, and Policy Evaluation 3. Implementing a Simple Markov Decision Process (MDP) 4. Value Iteration and Bellman Equation Demonstration 5. Implementing Q-Learning in a Gridworld Environment 6. Exploration vs Exploitation and Introduction to Deep Q-Network (DQN)	7

Elective course : INTRODUCTION TO ECONOMETRICS

Course Code: PBD-3901
No. of Credits: 04
Learning Hours: 60 hrs
Practical's to be conducted using R

Course Outcomes

- CO 1: The students will be able to understand how to undertake empirical research and analysis
- CO 2: The students will be able to appreciate the strengths and weaknesses of various econometric techniques
- CO 3: The student will be able to evaluate competing economic theories and alternative policies
- CO4 : The student will be able to understand the intricacies of various economic variables involved in a big data and learn to model them

Course Overview and Course Objectives

Unit 1) Analysis of panel data and Generalized Method of Moments
Unit 2) Simultaneous Equations System
Unit d) Cointegration
Unit e) Models in Econometrics

Unit No.	Name of Unit	No. of Lectures
Unit-1	**Analysis of panel data and Generalized Method of Moments**	
	Introduction to econometric data, Analysis of Panel Data, Generalized Method of Moments (GMM).	15
Unit-2	**Simultaneous Equations System**	
	Least Squares, Bias Problem, Estimation Method.	15
Unit-3	**Cointegration**	
	Concept, two variable model, Engle-Granger Method, Vector autoregressions (VAR), Vector error correlation model (VECM).	15
Unit-4	**Models in Econometrics**	
	ARCH/GARCH/SV models, some important generalizations like EGARCH & GJR models, ARCH –M models.	15

SUGGESTED BOOKS:

1. The Econometrics of Financial Markets: J. Campbell, A.Lo and C. Mackinlay
2. Econometric Analysis: William H. Greene
3. Introduction to Econometrics : *Jeffrey M. Wooldridge.*

Elective course : TIME SERIES ANALYSIS & FORECASTING

Course Code: PBD-3902
No. of Credits: 04
Learning Hours: 60 hrs
Practical's to be conducted using R/Python

- CO 1: Understand Stationary and Non-stationary Time series
- CO 2: Construct different models using time series analysis
- CO 3: Apply forecasting in Time series analysis
- CO4 : Understand and apply modeling in Seasonal Time series data
- CO 5: Do missing data analysis in time series data.

Unit No.	Name of Unit	No. of Lectures
Unit-1	**Basic Time series Models**	
	Stationary and Non-Stationary Time Series, AR,MA, ARMA and ARIMA models with illustrations, properties, estimation of parameters	15
Unit-2	**Missing Data Problem in Time Series.**	
	Examples and case studies	15
Unit-3	**Test of Stationarity and Forecast error**	
	Tests of Non-Stationarity – Unit Root tests, Forecasting, Smoothing, Minimum MSE Forecast, Forecast Error.	15
Unit-4	Modelling Seasonal Time Series	
	SARIMA, Holt-Winters, STL, LSTM	15

SUGGESTED BOOKS

1. Introduction to Statistical Time Series: W. A. Fuller

2. Introduction to Time Series Analysis: P. J. Brockwell and R. A. Davis

****

SEMESTER – IV

Internship based project
Course Code: PBD-4801
No. of Credits: 20
Duration of Project: 20 weeks

A real life project has to be undertaken at an industry for 20 weeks. Each student will have two supervisors: one from academic institution and one from the industry. The project shall involve handling data extensively and use of methodologies learnt during the course work to derive meaningful inferences. A final project report has to be submitted and an "open" presentation has to be made.

Project evaluation may be as follows.

Report from two supervisors: 200 marks (100 each)
Project report: 100 marks
Grand Viva-Voce : 100 marks
Presentation: 100 marks.
Total: 500 marks

****