Department
of
Computer Science and Engineering

B.Tech. (Minor Degree) Programme
in
DATA SCIENCE (CSD)

COURSE STRUCTURE & SYLLABUS

B.Tech. with Minor Degree Programme in Data Science

Course Structure (2021-22)

(Applicable for 2019-2023 Batch-R19 Syllabus)

III YEAR I & II SEMESTER

Year/Semester	Theory	Laboratory	Total Credits
III - I Semester	Introduction to Data Science (3 Hours, 3 Credits)	R Programming Laboratory (3 Hours, 1.5 Credits)	4.5
III - II Semester	Data Science Applications (3 Hours, 3 Credits)	Data Science Applications Laboratory (2 Hours, 1 Credit)	4
Total Credits			**8.5**

IV YEAR I & II SEMESTER

Year/Semester	Theory	Laboratory	Total Credits
IV - I Semester	(Either online through MOOCS or off-line Class) Data Wrangling and Visualization **OR** Big Data Analytics (3 Hours, 3 Credits)	(The corresponding Laboratory) Data Wrangling and Visualization **OR** Big Data Analytics	4.5
IV - II Semester	**Any one of the following subjects:** (which is not studied in regular B. Tech. course) 1. Exploratory Data Analysis 2. Mining Massive Databases 3. Social Network Analysis 4. Predictive Analytics 5. Web & Social Media Analytics 6. Video Analytics (3 Hours, 3 Credits)	--	3
IV-II Semester	Mini Project	--	2
Total Credits			**9.5**
Grand Total			**18**

B.Tech. Data Science (Minor) III Year I Sem.

INTRODUCTION TO DATA SCIENCE

L T P C
3 0 0 3

Course Objectives:

- Learn concepts, techniques and tools they need to deal with various facets of data science practice, including data collection and integration
- Understand the basic types of data and basic statistics
- Identify the importance of data reduction and data visualization techniques

Course Outcomes: After completion of the course, the student should be able to

- CO-1: Understand basic terms what Statistical Inference means. Identify probability distributions commonly used as foundations for statistical modeling. Fit a model to data
- CO-2: describe the data using various statistical measures
- CO-3: utilize R elements for data handling
- CO-4: perform data reduction and apply visualization techniques.

UNIT-I: Introduction

What is Data Science? - Big Data and Data Science hype – and getting past the hype - Datafication - Current landscape of perspectives - Statistical Inference - Populations and samples - Statistical modeling, probability distributions, fitting a model – Over fitting.

Basics of R: Introduction, R-Environment Setup, Programming with R, Basic Data Types.

UNIT-II: Data Types & Statistical Description

Types of Data: Attributes and Measurement, What is an Attribute? The Type of an Attribute, The Different Types of Attributes, Describing Attributes by the Number of Values, Asymmetric Attributes, Binary Attribute, Nominal Attributes, Ordinal Attributes, Numeric Attributes, Discrete versus Continuous Attributes.

Basic Statistical Descriptions of Data: Measuring the Central Tendency: Mean, Median, and Mode, Measuring the Dispersion of Data: Range, Quartiles, Variance, Standard Deviation, and Inter-quartile Range, Graphic Displays of Basic Statistical Descriptions of Data.

UNIT-III

Vectors: Creating and Naming Vectors, Vector Arithmetic, Vector sub setting,

Matrices: Creating and Naming Matrices, Matrix Sub setting, Arrays, Class.

Factors and Data Frames: Introduction to Factors: Factor Levels, Summarizing a Factor, Ordered Factors, Comparing Ordered Factors, Introduction to Data Frame, sub setting of Data Frames, Extending Data Frames, Sorting Data Frames.

Lists: Introduction, creating a List: Creating a Named List, Accessing List Elements, Manipulating List Elements, Merging Lists, Converting Lists to Vectors

UNIT-IV

Conditionals and Control Flow: Relational Operators, Relational Operators and Vectors, Logical Operators, Logical Operators and Vectors, Conditional Statements.

Iterative Programming in R: Introduction, While Loop, For Loop, Looping Over List.

Functions in R: Introduction, writing a Function in R, Nested Functions, Function Scoping, Recursion, Loading an R Package, Mathematical Functions in R.

UNIT-V:

Data Reduction: Overview of Data Reduction Strategies, Wavelet Transforms, Principal Components Analysis, Attribute Subset Selection, Regression and Log-Linear Models: Parametric Data Reduction, Histograms, Clustering, Sampling, Data Cube Aggregation.

Data Visualization: Pixel-Oriented Visualization Techniques, Geometric Projection Visualization Techniques, Icon-Based Visualization Techniques, Hierarchical Visualization Techniques, Visualizing Complex Data and Relations.

TEXT BOOKS:

1. Doing Data Science, Straight Talk from The Frontline. Cathy O'Neil and Rachel Schutt, O'Reilly, 2014
2. Jiawei Han, Micheline Kamber and Jian Pei. Data Mining: Concepts and Techniques, 3rd ed. The Morgan Kaufmann Series in Data Management Systems.
3. K G Srinivas, G M Siddesh, "Statistical programming in R", Oxford Publications.

REFERENCE BOOKS:

1. Introduction to Data Mining, Pang-Ning Tan, Vipin Kumar, Michael Steinbanch, Pearson Education.
2. Brain S. Everitt, "A Handbook of Statistical Analysis Using R", Second Edition, 4 LLC, 2014.
3. Dalgaard, Peter, "Introductory statistics with R", Springer Science & Business Media, 2008.
4. Paul Teetor, "R Cookbook", O'Reilly, 2011.

B.Tech. Data Science (Minor) III Year I Sem.

R PROGRAMMING LABORATORY

L T P C
0 0 3 1.5

1. R Environment setup: Installation of R and RStudio in Windows
2. Write R commands for
   i. Variable declaration and retrieving the value of the stored variables,
   ii. Write an R script with comments,
   iii. Type of a variable using class () Function.
3. Write R command to
   i. illustrate summation, subtraction, multiplication, and division operations on vectors using vectors.
   ii. Enumerate multiplication and division operations between matrices and vectors in R console
4. Write R command to
   i. Illustrate the usage of Vector sub setting& Matrix sub setting
   ii. Write a program to create an array of 3×3 matrixes with 3 rows and 3 columns.
   iii. Write a program to create a class, object, and function
5. Write a command in R console
   i. to create a tshirt_factor, which is ordered with levels 'S', 'M', and 'L'. Is it possible to identify from the examples discussed earlier, if blood type 'O' is greater or less than blood type 'A'?
   ii. Write the command in R console to create a new data frame containing the 'age' parameter from the existing data frame. Check if the result is a data frame or not. Also R commands for data frame functions cbind(), rbind(), sort()
6. Write R command for
   i. Create a list containing strings, numbers, vectors and logical values
   ii. To create a list containing a vector, a matrix, and a list. Also give names to the elements in the list and display the list also access the list elements
   iii. To add a new element at the end of the list and delete the element from the middle display the same
   iv. To create two lists, merge two lists. Convert the lists into vectors and perform addition on the two vectors. Display the resultant vector.
7. Write R command for
   i. logical operators—AND (&), OR (|) and NOT (!).
   ii. Conditional Statements
   iii. Create four vectors namely patient id, age, diabetes, and status. Put these four vectors into a data frame patient data and print the values using a for loop& While loop
   iv. Create a user-defined function to compute the square of an integer in R
   v. Create a user-defined function to compute the square of an integer in R
   vi. Recursion function for a) factorial of a number b) find nth Fibonacci number
8. Write R code for i) Illustrate Quick Sort ii) Illustrate Binary Search Tree
9. Write R command to
   i. illustrate Mathematical functions & I/O functions
   ii. Illustrate Naming of functions and sapply(), lapply(), tapply() &mapply()
10. Write R command for

i. Pie chart& 3D Pie Chart, Bar Chart to demonstrate the percentage conveyance of various ways for traveling to office such as walking, car, bus, cycle, and train

ii. Using a chart legend, show the percentage conveyance of various ways for traveling to office such as walking, car, bus, cycle, and train.

a. Walking is assigned red color, car – blue color, bus – yellow color, cycle – green color, and train – white color; all these values are assigned through cols and lbls variables and the legend function.

b. The fill parameter is used to assign colors to the legend.

c. Legend is added to the top-right side of the chart, by assigning

iii. Using box plots, Histogram, Line Graph, Multiple line graphs and scatter plot to demonstrate the relation between the cars speed and the distance taken to stop, Consider the parameters data and x Display the speed and dist parameter of Cars data set using x and data parameters

TEXT BOOK:

1. K G Srinivas, G M Siddesh, "Statistical programming in R", Oxford Publications.

Department of Computer Science and Engineering

B.Tech. with Minor program in Data Science

III Year-II Semester **L T P C**

DATA SCIENCE APPLICATIONS **3 0 0 3**

Course Objective: To give deep knowledge of data science and how it can be applied in various Fields to make the life easy.

Course Outcomes: After completion of course, students would:
1. To correlate the data science and solutions to modern problem.
2. To decide when to use which type of technique in data science.

UNIT - I
Data Science Applications in various domains, Challenges and opportunities, tools for data scientists Recommender systems – Introduction, methods, application, challenges.

UNIT - II
Time series data – stock market index movement forecasting.
Supply Chain Management – Real world case study in logistics

UNIT - III
Data Science in Education, Social media, Student Result Analysis, Facebook Data Analysis.

UNIT - IV
Data Science in Healthcare, Bioinformatics, Covid-19 data Analysis, Biological Data (DNA Sequencing Data)

UNIT - V
Case studies in data optimization using Python.

TEXT BOOKS:
1. Aakanksha Sharaff, G.K.Sinha , "Data Science and its applications ", CRC Press, 2021.
2. Q.A.Menon, S.A.Khoja, "Data Science: Theory, Analysis and Applications", CRC Press, 2020.

Department of Computer Science and Engineering

B.Tech. Data Science (Minor) III Year II Sem.

Data Science Applications Laboratory

L	T	P	C
0	0	2	1

Course Objective: To give deep knowledge of Data Science and how Data Science can be applied in various fields to make the life easy.

Course Outcomes: After completion of course, students would know how:

1. To correlate the Data Science solutions to modern problems.
2. To decide when to use which type of Data Science technique.

List of Experiments

1. Applications of Data Science in Health Care
2. Applications of Data Science in Stock Market Index Movement Forecasting
3. Applications of Data Science in Student Result Analysis
4. Applications of Data Science in Weather Fore Casting
5. Applications of Data Science in Facebook Data Analysis
6. Applications of Data Science in Covid 19 Data Analysis
7. Applications of Data Science in Biological Data Analysis
8. Applications of Data Science in Banking Data Analysis
9. Applications of Data Science in predictions of Vitamins in food
10. Applications of Data Science in DNA Sequence analysis.

B.Tech. Data Science (Minor) IV Year I Sem.

Data Wrangling and Visualization

L T P C
3 - - 3

Course Objectives:

1. Understand the basic concepts of data wrangling.
2. Learn the different data import methods.
3. Learn the concepts of web scraping
4. Understand the visualization process and visual representations of data.
5. Learn visualization techniques for various types of data.
6. Explore the visualization techniques for graphs, trees, Networks.

Course outcome:

1. Identify and execute the basic data format.
2. Perform the computations with Excel and pdf files.
3. Explore and analyze the Image and video data.
4. Understand the concepts web scraping.
5. Apply the visualization process for creating visual representations.

UNIT – I

Data Wrangling: Definition, Importance, How it can be performed?, Tasks, Tools. Pandas in python: Pandas Basics, Read the data from Machines as CSV Data, JSON Data, XML Data.

WORKING WITH EXCEL FILES AND PDFS: Getting Started with Parsing, Installing Python Packages, Parsing Excel Files, Programmatic Approaches to PDF Parsing, Converting PDF to Text-Parsing PDFs Using pdf miner. A Brief Introduction -Relational Databases: MySQL and Postgre SQL; Non-Relational Databases: No SQL; When to Use a Simple File or Data Storage.

UNIT -II

Data Clean up: Why Clean Data?, Data Cleanup Basics, Identifying missing Values in Data, Formatting Data, Finding Outliers, Finding Duplicates, Fuzzy Matching, RegEx Matching, Normalizing and Standardizing the Data, making data consistent, grouping data values into bins, and converting categorical variables into numerical quantitative variables, Saving the Data, Determining suitable Data Cleanup methods, Scripting the Cleanup, Testing with New Data

Data Exploration and Analysis: Exploring Data, Importing Data, Exploring Table Functions, Joining Numerous Datasets, Identifying mean, median, mode, quartile values, and pearson correlations, Identifying Outliers, Creating Groupings, Analyzing Data, Separating and Focusing the Data, Presenting Data, Presentation Tools, publishing the Data, Open Source Platforms.

UNIT – III

WEB SCRAPING: What to Scrape and How?, Analyzing a Web Page, Getting Pages, Reading a Web Page, Reading a Web Page with LXML, Xpath; Advanced Web Scraping: Browser-Based Parsing, Screen Reading with Selenium or Ghost.Py, Spidering the Web.

UNIT – IV

Visualization : What Is Visualization?, History of Visualization, Relationship between Visualization and Other fields, The Visualization Process, The Role of Cognition, Pseudocode Conventions, The Scatterplot, The Role of the User, Installing matplotlib, NumPy, and SciPy, Customizing matplotlib's parameters in code,

Plots: Defining plot types – bar, line, and stacked charts; Simple sine and cosine plot, Defining axis lengths and limits, Defining plot line styles, properties, and format strings; Setting ticks, labels, and grids; Adding a legend and annotations, Moving spines to the center, histograms, bar charts with error bars, pie charts, Plotting with filled areas, scatter plots with colored markers, heat maps.

UNIT – V

More Plots and Customizations: Setting the transparency and size of axis labels, Adding a shadow to the chart line, Adding a data table to the figure, Using subplots, Customizing grids, Creating contour plots, Filling an under-plot area, Drawing polar plots, Visualizing the file system tree using a polar bar.

3D Visualizations: Creating 3D bars, 3D histograms, Animating in matplotlib, Animating with OpenGL

Visualization Techniques: for Spatial Data, for Geospatial Data, for Time-Oriented Data, for Multivariate Data, Text and Document Visualization.

Text Books:

1. Jacqueline Kazil & Katharine Jarmul," Data Wrangling with Python", O'Reilly Media, Inc, 2016.
2. Dr. Tirthajyoti Sarkar, Shubhadeep," Data Wrangling with Python: Creating actionable data from raw sources", Packt Publishing Ltd, 2019.
3. Matthew Ward Georges Grinstein Daniel Keim , Interactive Data Visualization: Foundations, Techniques, and Applications. A K Peters, Ltd. Natick.
4. Igor Milovanović, Python Data Visualization:Cookbook,packet publishing, 2013.

REFERENCE BOOKS

1. Stefanie Molin," Hands-On Data Analysis with Pandas", Packt Publishing Ltd, 2019.
2. Allan Visochek," Practical Data Wrangling", Packt Publishing Ltd,2017
3. Tye Rattenbury, Joseph M. Hellerstein, Jeffrey Heer, Sean Kandel, Connor Carreras," Principles of Data Wrangling: Practical Techniques for Data Preparation", O'Reilly Media, Inc,2017.
4. Data Visualization: A Handbook for Data Drive by AndyKirk

B.Tech. Data Science (Minor) IV Year I Sem.

BIG DATA ANALYTICS

L T P C
3 - - 3

Course Objectives:

1. The purpose of this course is to provide the students with the knowledge of Big data Analyticsprinciples and techniques.
2. This course is also designed to give an exposure of the frontiers of Big data Analytics

Courses Outcomes:

1. Ability to explain the foundations, definitions, and challenges of Big Data and various Analyticaltools.
2. Ability to program using HADOOP and Map reduce, NOSQL
3. Ability to understand the importance of Big Data in Social Media and Mining.

UNIT - I

Introduction to Big Data: Big Data and its Importance – Four V's of Big Data – Drivers for Big Data –Introduction to Big Data Analytics – Big Data Analytics applications.

UNIT - II

Big Data Technologies: Hadoop's Parallel World – Data discovery – Open source technology for Big Data Analytics – cloud and Big Data –Predictive Analytics – Mobile Business Intelligence and Big Data

UNIT - III

Introduction Hadoop: Big Data – Apache Hadoop & Hadoop Eco System – Moving Data in and out ofHadoop – Understanding inputs and outputs of MapReduce - Data Serialization.

UNIT - IV

Hadoop Architecture: Hadoop: RDBMS Vs Hadoop, Hadoop Overview, Hadoop distributors, HDFS, HDFS Daemons, Anatomy of File Write and Read., Name Node, Secondary Name Node, and Data Node, HDFS Architecture, Hadoop Configuration, Map Reduce Framework, Role of HBase in Big Dataprocessing, HIVE, PIG.

UNIT - V

Data Analytics with R Machine Learning: Introduction, Supervised Learning, Unsupervised Learning, Collaborative Filtering, Social Media Analytics, Mobile Analytics, Big Data Analytics with BigR.

TEXT BOOKS:

1. Big Data Analytics, Seema Acharya, Subhasini Chellappan, Wiley 2015.
2. Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Business, Michael Minelli, Michehe Chambers, 1st Edition, Ambiga Dhiraj, Wiely CIO Series, 2013.
3. Hadoop: The Definitive Guide, Tom White, 3rd Edition, O"Reilly Media, 2012.

4. Big Data Analytics: Disruptive Technologies for Changing the Game, Arvind Sathi, 1st Edition, IBM Corporation, 2012.

REFERENCE BOOKS:

1. Big Data and Business Analytics, Jay Liebowitz, Auerbach Publications, CRC press (2013)
2. Using R to Unlock the Value of Big Data: Big Data Analytics with Oracle R Enterprise and Oracle R Connector for Hadoop, Tom Plunkett, Mark Hornick, McGraw-Hill/Osborne Media (2013), Oracle press.
3. Professional Hadoop Solutions, Boris lublinsky, Kevin t. Smith, Alexey Yakubovich, Wiley, ISBN: 9788126551071, 2015.
4. Understanding Big data, Chris Eaton, Dirk deroos et al. McGraw Hill, 2012.
5. Intelligent Data Analysis, Michael Berthold, David J. Hand, Springer, 2007.
6. Taming the Big Data Tidal Wave: Finding Opportunities in Huge Data Streams with AdvancedAnalytics, Bill Franks, 1st Edition, Wiley and SAS Business Series, 2012.

B.Tech. Data Science (Minor) IV Year I Sem.

DATA WRANGLING AND VISUALIZATION LAB

L T P C
- - 3 1.5

Course Objectives:

1. understand the different data importing method.
2. Make use of various data clean up techniques.
3. Understand the web wraping mechanisms.
4. Make more effective visualizations for data.
5. Understand how fundamental principles of design and human cognition inform effective visualizations.

Course Outcomes:

1. Perform Read and write operations on CSV,JSON and XML files
2. Process the Excel file using Pandas
3. Parse and Extract the Tables using Python library
4. Explore the web scraping in Python
5. Demonstrate knowledge of technical advances through active participation in life-long
6. Conduct research relevant data visualization topics
7. Use existing visualization tools and techniques to analyze basic datasets.

Week 1

1. Write a Python script to read each row from a given csv file and print a list of strings.
2. Write a Python program to read a given CSV file as a dictionary.

Week 2

3. Write a Python program to convert Python dictionary object (sort by key) to JSON data. Print the object members with indent level 4.

Week 3

4. Write the python script to Read the XML file
5. Write a Pandas program to import excel data (child labour and child marriage data.xlsx) into a Pandas data frame and process the following
   1. Get the data types of the given excel data.
   2. Display the last ten rows.
   3. Insert a column in the sixth position of the said excel sheet and fill it with NaN values

Week 4

6. Develop the python script to parse the pdf files using pdfminer.
7. Extract the Table from the child labour and child marriage data.xlsx using pdfables library

Week 5

8. Write a Python data wrangling scripts to insert the data into SQLite database
9. Develop the Python Shell Script to do the basic data cleanup on child labour and child marriage data.xlsx
   1. Check duplicates and missing data
   2. Eliminate Mismatches
   3. Cleans line breaks, spaces, and special characters

Week 6

10. Write a Python program to download and display the content of robot.txt for en.wikipedia.org

Week 7

11. Experiment Data Representation: chart types: categorical, hierarchical, relational, temporal & spatial;

Week 8

1. 2-D experiments: bar charts, Clustered bar charts, dot plots, connected dot plots, pictograms, proportional shape charts, bubble charts, radar charts, polar charts, Range chart, Box-and-whisker plots, univariate scatter plots, histograms word cloud, pie chart, waffle chart, stacked bar chart, back-to-back bar chart, tree map and all relevant 2-D charts.

Week 9

2. Experiment: surfaces, contours, hidden surfaces, pm3d coloring, 3Dmapping;

Week 10

1. Program on multi-dimensional data visualization, manifold visualization;

Text books:

1. Jacqueline Kazil & Katharine Jarmul," Data Wrangling with Python", O'Reilly Media, Inc,2016.
2. Dr. Tirthajyoti Sarkar, Shubhadeep," Data Wrangling with Python: Creating actionable data from raw sources", Packt Publishing Ltd,2019.
3. Andy Kirk, Data Visualization A Handbook for Data Driven Design, Sage Publications,2016
4. Philipp K. Janert, Gnuplot in Action, Understanding Data with Graphs, Manning Publications, 2010.
5. Sinan ozdemmir , "Principles of Data Science", PacketPublishers-2016

B.Tech. Data Science (Minor) IV Year I Sem.

BIG DATA ANALYTICS LAB

L	T	P	C
-	-	3	1.5

Course Objectives:

1. The purpose of this course is to provide the students with the knowledge of Big data Analyticsprinciples and techniques.
2. This course is also designed to give an exposure of the frontiers of Big data Analytics

Course Outcomes:

1. Use Excel as an Analytical tool and visualization tool.
2. Ability to program using HADOOP and Map reduce.
3. Ability to perform data analytics using ML in R.
4. Use cassandra to perform social media analytics.

List of Experiments:

1. Implement a simple map-reduce job that builds an inverted index on the set of inputdocuments (Hadoop)
2. Process big data in HBase
3. Store and retrieve data in Pig
4. Perform Social media analysis using cassandra
5. Buyer event analytics using Cassandra on suitable product sales data.
6. Using Power Pivot (Excel) Perform the following on any dataset
   a) Big Data Analytics
   b) Big Data Charting
7. Use R-Project to carry out statistical analysis of big data
8. Use R-Project for data visualization of social media data

TEXT BOOKS:

1. Big Data Analytics, Seema Acharya, Subhashini Chellappan, Wiley 2015.
2. Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Business, Michael Minelli, Michehe Chambers, 1st Edition, Ambiga Dhiraj, Wiely CIO Series, 2013.
3. Hadoop: The Definitive Guide, Tom White, 3rd Edition, O"Reilly Media, 2012.
4. Big Data Analytics: Disruptive Technologies for Changing the Game, Arvind Sathi, 1st Edition, IBM Corporation, 2012.

REFERENCE BOOKS:

1. Big Data and Business Analytics, Jay Liebowitz, Auerbach Publications, CRC press (2013).
2. Using R to Unlock the Value of Big Data: Big Data Analytics with Oracle R Enterprise and Oracle R Connector for Hadoop, Tom Plunkett, Mark Hornick, McGraw-Hill/Osborne Media (2013), Oracle press.

3. Professional Hadoop Solutions, Boris lublinsky, Kevin t. Smith, Alexey Yakubovich, Wiley, ISBN: 9788126551071, 2015.
4. Understanding Big data, Chris Eaton, Dirk deroos et al., McGraw Hill, 2012.
5. Intelligent Data Analysis, Michael Berthold, David J. Hand, Springer, 2007.
6. Taming the Big Data Tidal Wave: Finding Opportunities in Huge Data Streams with AdvancedAnalytics, Bill Franks, 1st Edition, Wiley and SAS Business Series, 2012.

B.Tech. Data Science (Minor) IV Year II Sem.

Exploratory Data Analysis

L T P C
3 - - 3

Course Objectives:

1. This course introduces the methods for data preparation and data understanding.
2. It covers essential exploratory techniques for understanding multivariate data by summarizingit through statistical methods and graphical methods.
3. Supports to Summarize the insurers use of predictive analytics, data science and Data Visualization.

Course Outcomes:

1. Handle missing data in the real-world data sets by choosing appropriate methods.
2. Summarize the data using basic statistics. Visualize the data using basic graphs and plots.
3. Identify the outliers if any in the data set.
4. Choose appropriate feature selection and dimensionality reduction.
5. Techniques for handling multi-dimensional data.

UNIT - I:

Introduction to Exploratory Data Analysis: Data Analytics lifecycle, Exploratory Data Analysis (EDA)– Definition, Motivation, Steps in data exploration, The basic data types Data Type Portability.

UNIT - II:

Preprocessing - Traditional Methods and Maximum Likelihood Estimation: Introduction to Missing data, Traditional methods for dealing with missing data, Maximum Likelihood Estimation – Basics, Missing data handling, Improving the accuracy of analysis. **Preprocessing Bayesian Estimation:** Introduction to Bayesian Estimation, Multiple Imputation-Imputation Phase, Analysis and Pooling Phase, Practical Issues in Multiple Imputation, Models for Missing Notation Random Data.

UNIT - III:

Data Summarization & Visualization: Statistical data elaboration, 1-D Statistical data analysis, 2-D Statistical data Analysis, N-D Statistical data analysis.

UNIT - IV:

Outlier Analysis: Introduction, Extreme Value Analysis, Clustering based, Distance Based and Density Based outlier analysis, Outlier Detection in Categorical Data. **Feature Subset Selection:** Feature selection algorithms: filter methods, wrapper methods and embedded methods, Forward selection backward elimination, Relief, greedy selection, genetic algorithms for features selection.

UNIT - V

Dimensionality Reduction: Introduction, Principal Component Analysis (PCA), Kernel PCA, Canonical Correlation Analysis, Factor Analysis, Multidimensional scaling, Correspondence

Analysis.

TEXT BOOKS:

1. Making sense of Data: A practical Guide to Exploratory Data Analysis and Data Mining, byGlenn J. Myatt.

REFERENCE BOOKS:

1. Charu C. Aggarwal, "Data Mining The Text book", Springer, 2015.
2. Craig K. Enders, "Applied Missing Data Analysis", The Guilford Press, 2010.
3. Inge Koch, "Analysis of Multivariate and High dimensional data", Cambridge University Press,2014.
4. Michael Jambu, "Exploratory and multivariate data analysis", Academic Press Inc., 1990.
5. Charu C. Aggarwal, "Data Classification Algorithms and Applications", CRC press, 2015.

B.Tech. Data Science (Minor) IV Year II Sem.

Mining Massive Databases

L T P C
3 - - 3

Prerequisites: Students should be familiar with Data mining, algorithms, basic probability theory and Discrete math.

Course Objectives:

1. This course will cover practical algorithms for solving key problems in mining of massive datasets.
2. This course focuses on parallel algorithmic techniques that are used for large datasets.
3. This course will cover stream processing algorithms for data streams that arrive constantly, page ranking algorithms for web search, and online advertisement systems that are studied indetail.

Course Outcomes:

1. Handle massive data using MapReduce.
2. Develop and implement algorithms for massive data sets and methodologies in the context ofdata mining.
3. Understand the algorithms for extracting models and information from large datasets
4. Develop recommendation systems.
5. Gain experience in matching various algorithms for particular classes of problems.

UNIT - I

Data Mining-Introduction-Definition of Data Mining-Statistical Limits on Data Mining, **MapReduce and the New Software Stack**-Distributed File Systems, MapReduce, Algorithms Using MapReduce.

UNIT - II

Similarity Search: Finding Similar Items-Applications of Near-Neighbor Search, Shingling of Documents, Similarity-Preserving Summaries of Sets, Distance Measures. **Streaming Data:** Mining Data Streams-The Stream Data Model, Sampling Data in a Stream, Filtering Streams.

UNIT - III

Link Analysis-PageRank, Efficient Computation of PageRank, Link Spam. **Frequent Itemsets -** Handling Larger Datasets in Main Memory, Limited-Pass Algorithms, Counting Frequent Items in a Stream. **Clustering**-The CURE Algorithm, Clustering in Non-Euclidean Spaces, Clustering for Streamsand Parallelism.

UNIT - IV

Advertising on the Web-Issues in On-Line Advertising, On-Line Algorithms, The Matching Problem, The Adwords Problem, Adwords Implementation. Recommendation Systems - A Model for Recommendation Systems, Content-Based Recommendations, Collaborative Filtering, DimensionalityReduction, The NetFlix Challenge.

UNIT - V

Mining Social-Network Graphs-Social Networks as Graphs, Clustering of Social-Network Graphs, Partitioning of Graphs, Simrank, Counting Triangles.

TEXT BOOKS:

1. Jure Leskovec, Anand Rajaraman, Jeff Ullman, Mining of Massive Datasets, 3rd Edition.

REFERENCE BOOKS:

1. Jiawei Han & Micheline Kamber, Data Mining – Concepts and Techniques 3rd Edition Elsevier.
2. Margaret H Dunham, Data Mining Introductory and Advanced topics, PEA.
3. Ian H. Witten and Eibe Frank, Data Mining: Practical Machine Learning Tools and Techniques,Morgan Kaufmann.

B.Tech. Data Science (Minor) IV Year II Sem.

Social Network Analysis

L T P C
3 - - 3

Course Objectives:

The student should be made to:

1. Understand the concept of semantic web and related applications.
2. Learn knowledge representation using ontology.
3. Understand human behaviour in social web and related communities.
4. Learn visualization of social networks.

Course Outcomes:

Upon completion of the course, the student should be able to:

1. Develop semantic web related applications.
2. Represent knowledge using ontology.
3. Predict human behaviour in social web and related communities.
4. Visualize social networks.

UNIT I

Introduction: Introduction to Semantic Web: Limitations of current Web - Development of Semantic Web -Emergence of the Social Web - Social Network analysis: Development of Social Network Analysis - Key concepts and measures in network analysis - Electronic sources for network analysis: Electronic discussion networks, Blogs and online communities - Web-based networks - Applications of Social Network Analysis.

UNIT II

Modelling, Aggregating And Knowledge Representation: Ontology and their role in the Semantic Web: Ontology-based knowledge Representation - Ontology languages for the Semantic Web: Resource Description Framework - Web Ontology Language - Modelling and aggregating social network data: State-of-the-art in network data representation - Ontological representation of social individuals - Ontological representation of social relationships - Aggregating and reasoning with social network data - Advanced representations.

UNIT III

Extraction And Mining Communities In Web Social Networks: Extracting evolution of Web Community from a Series of Web Archive - Detecting communities in social networks - Definition of community - Evaluating communities - Methods for community detection and mining - Applications of community mining algorithms - Tools for detecting communities social network infrastructures and communities - Decentralized online social networks - Multi- Relational characterization of dynamic social network communities.

UNIT IV

Predicting Human Behaviour And Privacy Issues: Understanding and predicting human behaviour for social communities - User data management - Inference and Distribution - Enabling new human experiences - Reality mining - Context - Awareness - Privacy in online social networks - Trust in online environment - Trust models based on subjective logic - Trust network analysis - Trust

transitivity analysis - Combining trust and reputation - Trust derivation based on trust comparisons - Attack spectrum and countermeasures.

UNIT V

Visualization And Applications Of Social Networks: Graph theory - Centrality - Clustering - Node-Edge Diagrams - Matrix representation - Visualizing online social networks, Visualizing social networks with matrix-based representations - Matrix and Node-Link Diagrams - Hybrid representations - Applications - Cover networks - Community welfare - Collaboration networks - Co-Citation networks.

TEXT BOOKS:

1. Peter Mika, "Social Networks and the Semantic Web", First Edition, Springer 2007.

2. Borko Furht, "Handbook of Social Network Technologies and Applications", 1st Edition, Springer, 2010.

REFERENCES:

1. Guandong Xu ,Yanchun Zhang and Lin Li, "Web Mining and Social Networking – Techniques and applications", First Edition Springer, 2011.

2. Dion Goh and Schubert Foo, "Social information Retrieval Systems: Emerging Technologies and Applications for Searching the Web Effectively", IGI Global Snippet, 2008.

3. Max Chevalier, Christine Julien and Chantal Soulé-Dupuy, "Collaborative and Social Information Retrieval and Access: Techniques for Improved user Modelling", IGI Global Snippet, 2009.

4. John G. Breslin, Alexander Passant and Stefan Decker, "The Social Semantic Web", Springer, 2009.

B.Tech. Data Science (Minor) IV Year II Sem.

Predictive Analysis

L T P C
3 - - 3

Course Objectives: The course serves to advance and refine expertise on theories, approaches and techniques related to prediction and forecasting.

Course Outcomes

1. Understand prediction-related principles, theories and approaches.
2. Learn model assessment and validation.
3. Understand the basics of predictive techniques and statistical approaches.
4. Analyze supervised and unsupervised algorithms.

UNIT - I

Linear Methods for Regression and Classification: Overview of supervised learning, Linear regression models and least squares, Multiple regression, Multiple outputs, Subset selection, Ridge regression, Lasso regression, Linear Discriminant Analysis, Logistic regression, Perceptron learning algorithm.

UNIT - II

Model Assessment and Selection: Bias, Variance, and model complexity, Bias-variance trade off, Optimism of the training error rate, Estimate of In-sample prediction error, Effective number of parameters, Bayesian approach and BIC, Cross- validation, Boot strap methods, conditional or expected test error.

UNIT - III

Additive Models, Trees, and Boosting: Generalized additive models, Regression and classification trees, Boosting methods-exponential loss and AdaBoost, Numerical Optimization via gradient boosting,Examples (Spam data, California housing, New Zealand fish, Demographic data).

UNIT - IV

Neural Networks (NN), Support Vector Machines (SVM), and K-nearest Neighbor: Fitting neural networks, Back propagation, Issues in training NN, SVM for classification, Reproducing Kernels, SVM for regression, K-nearest – Neighbour classifiers (Image Scene Classification).

UNIT - V

Unsupervised Learning and Random forests: Association rules, Cluster analysis, Principal Components, Random forests and analysis.

TEXT BOOK:

1. Trevor Hastie, Robert Tibshirani, Jerome Friedman, The Elements of Statistical Learning-DataMining, Inference, and Prediction, Second Edition, Springer Verlag, 2009.

REFERENCE BOOKS:

1. C.M.Bishop –Pattern Recognition and Machine Learning, Springer, 2006.
2. L. Wasserman-All of statistics.
3. Gareth James. Daniela Witten. Trevor Hastie Robert Tibshirani. An Introduction to StatisticalLearning with Applications in R.

B.Tech. Data Science (Minor) IV Year II Sem.

Web & Social Media Analytics

L T P C
3 - - 3

Course Objectives: Exposure to various web and social media analytic techniques.

Course Outcomes:

1. Knowledge on decision support systems.
2. Apply natural language processing concepts on text analytics.
3. Understand sentiment analysis.
4. Knowledge on search engine optimization and web analytics.

UNIT - I

An Overview of Business Intelligence, Analytics, and Decision Support: Analytics to Manage a Vaccine Supply Chain Effectively and Safely, Changing Business Environments and Computerized Decision Support, Information Systems Support for Decision Making, The Concept of Decision Support Systems (DSS), Business Analytics Overview, Brief Introduction to Big Data Analytics.

UNIT - II

Text Analytics and Text Mining: Machine Versus Men on Jeopardy!: The Story of Watson, Text Analytics and Text Mining Concepts and Definitions, Natural Language Processing, Text Mining Applications, Text Mining Process, Text Mining Tools.

UNIT - III

Sentiment Analysis: Sentiment Analysis Overview, Sentiment Analysis Applications, Sentiment Analysis Process, Sentiment Analysis and Speech Analytics.

UNIT - IV

Web Analytics, Web Mining: Security First Insurance Deepens Connection with Policyholders, Web Mining Overview, Web Content and Web Structure Mining, Search Engines, Search Engine Optimization, Web Usage Mining (Web Analytics), Web Analytics Maturity Model and Web Analytics Tools.

UNIT - V

Social Analytics and Social Network Analysis: Social Analytics and Social Network Analysis, SocialMedia Definitions and Concepts, Social Media Analytics.

Prescriptive Analytics - Optimization and Multi-Criteria Systems: Multiple Goals, Sensitivity Analysis, What-If Analysis, and Goal Seeking.

TEXT BOOK:

1. Ramesh Sharda, Dursun Delen, Efraim Turban, BUSINESS INTELLIGENCE AND ANALYTICS: SYSTEMS FOR DECISION SUPPORT, Pearson Education.

REFERENCE BOOKS:

1. Rajiv Sabherwal, Irma Becerra-Fernandez," Business Intelligence – Practice, Technologies andManagement", John Wiley 2011.
2. Lariss T. Moss, ShakuAtre, "Business Intelligence Roadmap", Addison-Wesley It Service.
3. Yuli Vasiliev, "Oracle Business Intelligence: The Condensed Guide to Analysis and Reporting",SPD Shroff, 2012.

B.Tech. Data Science (Minor) IV Year II Sem.

Video Analytics

L T P C
3 - - 3

Course Objectives: To acquire the knowledge of extracting information from surveillance videos, understand the models used for recognition of objects, humans in videos and perform gait analysis.

Course Outcomes:

1. Understand the basics of video- signals and systems.
2. Able to estimate motion in a video.
3. Able to detect the objects and track them.
4. Recognize activity and analyze behaviour.
5. Evaluate face recognition technologies.

UNIT - I

INTRODUCTION Multidimensional signals and systems: signals, transforms, systems, sampling theorem. Digital Images and Video: human visual system and color, digital video, 3D video, digital-videoapplications, image and video quality.

UNIT - II

MOTION ESTIMATION Image formation, motion models, 2D apparent motion estimation, differential methods, matching methods, non-linear optimization methods, transform domain methods, 3D motion and structure estimation.

UNIT - III

VIDEO ANALYTICS Introduction- Video Basics - Fundamentals for Video Surveillance- Scene Artifacts- Object Detection and Tracking: Adaptive Background Modelling and Subtraction- PedestrianDetection and Tracking Vehicle Detection and Tracking- Articulated Human Motion Tracking in Low- Dimensional Latent Spaces.

UNIT - IV

BEHAVIORAL ANALYSIS & ACTIVITY RECOGNITION Event Modelling- Behavioral Analysis- Human Activity Recognition-Complex Activity Recognition Activity modelling using 3D shape, Video summarization, shape-based activity models- Suspicious Activity Detection.

UNIT - V

HUMAN FACE RECOGNITION & GAIT ANALYSIS Introduction: Overview of Recognition algorithms – Human Recognition using Face: Face Recognition from still images, Face Recognition from video, Evaluation of Face Recognition Technologies- Human Recognition using gait: HMM Framework for Gait Recognition, View Invariant Gait Recognition, Role of Shape and Dynamics in Gait Recognition.

TEXT BOOKS:

1. Murat Tekalp, “Digital Video Processing”, second edition, Pearson, 2015

2. Rama Chellappa, Amit K. Roy-Chowdhury, Kevin Zhou. S, “Recognition of Humans and their Activities using Video”, Morgan & Claypool Publishers, 2005.
3. Yunqian Ma, Gang Qian, “Intelligent Video Surveillance: Systems and Technology”, CRC Press (Taylor and Francis Group), 2009.