HONOURS DEGREE – DATA SCIENCE
List of Courses

Course Code	SEMESTER	Course Name	L	T	P	C
21IT4S01	IV	FUNDAMENTALS OF DATA SCIENCE	3	0	0	3
21IT5S02	V	DATA MINING USING R	3	0	1	4
21IT6S03	VI	DATA VISUALIZATION FOR ENGINEERS	3	0	1	4
21IT7S04	VII	BUSINESS INTELLIGENCE AND ANALYTICS	3	0	0	3
21IT8S05	VIII	CREATIVE AND INNOVATIVE PROJECT	0	0	8	4

21IT4S01 FUNDAMENTALS OF DATA SCIENCE L T P C

3 0 0 3

Course Objective:

The objective of this course is

- To provide overview on mathematical principles and techniques needed for modern data analysis.
- To introduce key conceptual techniques for data analysis such as linear regression, Principal component analysis, Bayesian inferences etc.,
- To understand the life cycle of Data Science and computational environments for data scientists using Python.
- To describe the fundamentals for exploring and manipulating data with Python

PREREQUISITE: Fundamental of Mathematics.

UNIT I PROBABILITY AND LINEAR ALGEBRA 9

Conditional Probability and independence – Density functions – Expected value – Variance –conditional distribution- Bayes' rule- Bayesian Inference – Convergence and sampling : Estimation – Concentration of measure - Linear algebra review – Vectors – Matrices – Norms – Linear Independence – Rank – Inverse – Orthogonality.

UNIT II APPLIED MATHEMATICS FOR DATA SCIENCE 9

Metrics – Lp distances - Mahalanobis distance – Cosine and angular distance – KL divergence – Distances for sets and strings – Similarities – Linear regression: Simple linear regression, Multiple explanatory variables, Polynomial regression, Cross validation, Regularized regression, Principal Component Analysis, Markov Chains

UNIT III INTRODUCTION TO DATA SCIENCE 9

Introduction to Data Science and its importance - Data Science and Big data-, The life cycle of Data

Science- The Art of Data Science - Work with data – data Cleaning, data Munging, data manipulation. Establishing computational environments for data scientists using Python with IPython and Jupyter.

UNIT IV DATA EXPLORATION USING NUMPY 9

Understanding Data Types in Python - The Basics of NumPy Arrays - Computation on NumPy Arrays: Universal Functions - Aggregations: Min, Max, and Everything in Between Computation on Arrays: Broadcasting-Comparisons, Masks, and Boolean Logic Fancy Indexing-Sorting Arrays.

UNIT V DATA MANIPULATION USING PANDAS 9

Installing and Using Pandas, Introducing Pandas Objects,Data Indexing and Selection. Operating on Data in Pandas,Handling Missing Data, Hierarchical Indexing Combining Datasets: Concat and Append, Combining Datasets: Merge and Join. Aggregation and Grouping, Pivot Tables, Vectorized String Operations, Working with Time Series.

Total : 45 hours

Text Book:

1. Jeff M. Philips, “Mathematical Foundations for Data Analysis”,Springer series in data sciences, Revised edition, 2021
2. Python Data Science Handbook-Essential Tools for Working with Data, Jake Vander Plas, O'Reilly Media,2016.
3. Data Science from Scratch: First Principles with Python, Joel Grus, O'Reilly,2015

Reference books

1. Python for Data Analysis, Wes Mckinney, O'Reilly Media,2013.
2. Field Cady, “Data Science Hand Book”, John Wiley & Sons,2017.
3. Fundamentals of Data Science, Samuel Burns, Amazon KDP printing and Publishing,2019.
4. Doing Data Science, Straight Talk From The Frontline, Cathy O'Neil and Rachel Schutt. O'Reilly.2014.
5. Tony Ojeda, Sean Patrick Murphy, Benjamin Bengfort, AbhijitDasgupta, “Practical Data Science Cookbook”, Packt Publishing Ltd.,2014.
6. Nathan Yau, “Visualize This: The Flowing Data Guide to Design, Visualization, and Statistics”, Wiley,2011.
7. ShaiVaingast, “Beginning Python Visualization Crafting Visual Transformation Scripts”, Apress, 2nedition, 2014.

Web references:

- http://www.dataquest.io/course/pandas-fundamentals/
- https://onlinecourses.nptel.ac.in/noc18_cs28/
- https://pandas.pydata.org/pandas-docs/stable/reference/general_functions.html
- https://www.guru99.com/data-science-tutorial.html

COURSE OUTCOMES:

At the end of the course, the students will be able to

- **CO1** – Understand application of mathematics for data analysis and machine learning
- **CO2 –** To learn the probability distributions and density estimations to perform analysis of various kinds of data
- **CO 3 –** Identify various phases involved in the life cycle of Data Science
- **CO 4 –** Preprocess and manage the data for efficient storage and manipulation in Python
- **CO 5 –** Realize the various data analytics techniques for labeled/columnar Data using Python Pandas
- **CO 6 –** Explore a flexible range of data visualizations approaches inPython.

PO vs CO MAPPING

CO. No	PO_a	PO_b	PO_c	PO_d	PO_e	PO_f	PO_g	PO_h	PO_i	PO_j	PO_k	PO_l
	2		2						3			
	3	3	2		3	3	3					2
	2	2	3				2					
	3		2	2					2			
	3	3	2	3	2	2	2				3	2
	3											

1→Low 2→Medium 3→High

21IT5S02 DATA MINING USING R **L T P C**

3 0 1 4

OBJECTIVES:

To impart knowledge on

1. Fundamentals of data mining and Basic R programming
2. Understanding classification and regression techniques and applying using R
3. Implementation and visualization of clustering & outliers in R
4. Prediction based on association rules and time series analysis in R
5. Explore R for various applications.

PREREQUISITE: Fundamentals of Data science and Probability and statistics

UNIT I DATA MINING FUNDAMENTALS AND R BASICS 6

Introduction to Data Mining – Types of Data – Architecture – Knowledge Discovery Process - Basics of R – Working with Datasets in R – Data Import and Export – Save and Load- Data in Different Formats - Data Types – Vectors & operations – Matrices – Arrays – Factors & operations – Data Frames – Subsetting of Data Frames – List – Data Exploration and Visualization

UNIT II CLASSIFICATION AND REGRESSION 6

Supervised Learning – Classification – Decision Tress – Working with party and rpart module – Random Forest –Regression – Linear Regression – Logistic Regression – Non Linear Regression

UNIT III CLUSTERING AND OUTLIER DETECTION 6

Unsupervised Learning – Partition based methods : K-Means Clustering – K-Medoids Clustering – Hierarchical Clustering – Density-based Clustering – Outlier Detection – Univariate Outlier Detection – Detect by Clustering – Comparative analysis

UNIT IV TIME SERIES AND ASSOCIATION RULE 6

Time Series Data in R – Decomposition – Time Series Forecasting – Time Series Clustering – Time Series Classification – Association Rule Mining – Removing Redundancy – Interpreting Rules – Visualizing Association Rules

UNIT V TEXT MINING & SOCIAL NETWORK ANALYSIS 6

Text Mining – Applications in R – Social Network Analysis – Network of Terms – Network of Tweets – Two-Mode Network – Analysis and Forecasting of House Price Indices - Customer Response Prediction and Profit Optimization

TOTAL HOURS: 30

LIST OF EXPERIMENTS

1. Data Exploration with R
2. Visualizing data using ggplot
3. Prediction using linear regression in R

4. Prediction using logistic regression
5. Implement k-means clustering in R
6. Implementation of Decision tree classifier in R
7. Naïve Bayes classifier implementation in R
8. Implement Association rule mining in R
9. Implement Time series analysis in R

30 Hours

Text Book:

1. Yanchang Zhao, "R and Data Mining: Examples and Case Studies", Academic Press, First Edition,2013
2. Jiawei Han and MichelineKamber, "Data Mining Concepts and Techniques", Elsevier - Morgan Kaufmann Publisher, Second Edition, 2012.
3. Thomas Mailund - "Beginning Data Science in R – Data Analysis, Visualization and modeling for data scientist", Springer, 2017.

Reference Book:

1. K.G.Srinivasa, G M Siddesh, Chetan Shetty, "Statistical Programming in R", Oxford University Press, New Delhi, 2017
2. John Chambers, "Software for Data Analysis: Programming with R ", Springer; 1st ed. 2008. , 2nd printing 2009 edition
3. Thomas Lumley," Complex Surveys: A Guide to Analysis Using R", Wiley Series in survey methodology,2010
4. Nicholas J. Horton, Ken Kleinman," Using R and RStudio for Data Management, Statistical Analysis, and Graphics" , CRC Press, Second edition,2015
5. John Maindonald, W. John Braun, "Data Analysis and Graphics Using R: An Example-Based Approach", University Press, Cambridge, Third edition,2010

Course Outcome:

At the end of the course, the students will be able to

- **CO1 –** Know the knowledge discovery mechanism and basic concepts in data mining
- **CO2 –** Carry out basic operations and perform import &export data using R
- **CO3 –** Understand and Evaluate supervised learning techniques inR
- **CO4 –** Use R to perform clustering and to detectoutliers
- **CO5 –** Explore data analysis for time series and build associationrules
- **CO5 –** Apply R for text mining and otherapplications

PO vs CO MAPPING

CO. No	PO_a	PO_b	PO_c	PO_d	PO_e	PO_f	PO_g	PO_h	PO_i	PO_j	PO_k	PO_l
CO1	3	2	2						2			
CO2	3	3	2		3	3						3
CO3		2	2				3					
CO4	3			2								
CO5		2	2	3	2	2	2					2
CO6	3											

1→Low 2→Medium 3→High

21IT6S03 DATA VISUALIZATION FOR ENGINEERS L T P C

3 0 1 4

OBJECTIVES:

The objective of this course is to enable the students to

- Inspect and interpret the engineering data and preparing meaningful and aesthetically pleasing scientific reports
- Understand data representations and mappings in order to produce sensible results
- Use their perception to better understand this data
- Understand data distributions, associations and time series

PREREQUISITE:

- Data Mining with R, Data Analysis

UNIT I INTRODUCTION TO VISUALIZATION 6

Visualizing Data-Mapping Data onto Aesthetics, Aesthetics and Types of Data, Scales Map Data Values onto Aesthetics, Coordinate Systems and Axes- Cartesian Coordinates, Nonlinear Axes, Coordinate Systems with Curved Axes, Color Scales-Color as a Tool to Distinguish, Color to Represent DataValues, Color as a Tool to Highlight, Directory of Visualizations-Amounts, Distributions, Proportions, x–y relationships, Geospatial Data

UNIT II VISUALIZING DISTRIBUTIONS 6

Visualizing Amounts-Bar Plots, Grouped and Stacked Bars, Dot Plots and Heat maps, Visualizing Distributions: Histograms and Density Plots-Visualizing a Single Distribution, Visualizing Multiple Distributions at the Same Time, Visualizing Distributions: Empirical Cumulative Distribution Functions and Q-Q Plots-Empirical Cumulative Distribution Functions, Highly Skewed Distributions, Quantile-Quantile Plots, Visualizing Many Distributions at Once-Visualizing Distributions Along the Vertical

Axis, Visualizing Distributions Along the Horizontal Axis

UNIT III VISUALIZING PROPORTIONS AND ASSOCIATIONS 6

Visualizing Proportions-A Case for Pie Charts, A Case for Side-by-Side Bars, A Case for Stacked Bars and Stacked Densities, Visualizing Proportions Separately as Parts of the Total ,Visualizing Nested Proportions- Nested Proportions Gone Wrong, Mosaic Plots and Treemaps, Nested Pies ,Parallel Sets. Visualizing Associations: Among Two or More Quantitative Variables-Scatterplots, Correlograms, Dimension Reduction, Paired Data.

UNIT IV TIME SERIES AND FORECASTING 6

Visualizing Time Series and Other Functions of an Independent Variable-Individual Time Series , Multiple Time Series and Dose–Response Curves, Time Series of Two or More Response Variables, Visualizing Trends-Smoothing, Showing Trends with a Defined Functional Form, Detrending and Time-Series Decomposition , Case study on weather forecasting data

UNIT V VISUALIZATION FOR ENGINEERING APPLICATIONS 6

Real time application development: Visualization for control engineering and predictive maintenance of machines – Construction data management through geo-spatial data visualization – Pollution control by visualizing air quality data – Stock Market Trend Prediction through time series analysis – Disaster management by visualizing associations.

LIST OF EXPERIMENTS

1. Histogram and Bar charts using R
2. Create different scatter plots for variables in any dataset
3. Enhancing Aesthetics with color scales
4. Illustrations for Heat maps and correlograms
5. Implementation of time series visualization
6. Visualizing associations and proportions
7. Generating 3D graphs
8. Visualizing geographic data with ggmap
9. Visualization of forecasting and trend analysis
10. Case study on Business data analysis and visualization

TOTAL HOURS: 60

Text Book:

1. Claus O.Wilke, “Fundamentals of Data visualization”, O. Reilly Media, First Edition, march 2019
2. Eric Pimpler, “Data Visualization and Exploration with R”, Geospatial Training Services, First edition, 2017

Reference books:

1. Tony Fischetti, Brett Lantz, R: Data Analysis and Visualization,O’Reilly ,2016

2. Robert I. Kabacoff ,"R in Action: Data Analysis and Graphics with R", Manning Publications, Second Edition, 2015
3. Nicholas J.Horton and Ken Kleinman,"Using R and R Studio for Data Management, Statistical analysis and Graphics", CRC Press, Taylor and Francis Group, Second Edition 2015

Web references:

1. https://www.analyticsvidhya.com/blog/2015/07/guide-data-visualization-r/
2. https://www.kdnuggets.com/2018/06/7-simple-data-visualizations-should-know-r.html

Course Outcomes

At the end of the course, students will be able to

- **CO1 -** Be familiar with key concepts, principles and methods in data visualization
- **CO2 –** Understand the value of visualization, specific techniques in information visualization and scientific visualization
- **CO3 -** Visualize the data in engineering applications and advertently make visual choices
- **CO4 -** Visualize data distributions and proportions
- **CO5 -** Understand trend prediction and uncertainties
- **CO6 -** Develop skills to both design and critique visualization

PO vs CO MAPPING

CO. No	PO_a	PO_b	PO_c	PO_d	PO_e	PO_f	PO_g	PO_h	PO_i	PO_j	PO_k	PO_l
CO1	2	2	2	2					2			
CO2	3	3	2		3	3	3					3
CO3	3	2	2	3								
CO4	3		3	2					3			
CO5		2	2	3	2	2	2				2	2
CO6												

1→Low 2→Medium 3→High

21IT7S04 BUSINESS INTELLIGENCE AND ANALYTICS L T P C

3 0 0 3

OBJECTIVES:

The student should be made to:

1. Be exposed with the basic rudiments of business intelligence system.
2. To understand the modeling aspects behind Business Intelligence.
3. To understand the business intelligence life cycle and the techniques used in it.
4. Be exposed with different data analysis tools and techniques.

PREREQUISITE:

- Fundamentals of Data Science

UNIT I BUSINESS INTELLIGENCE 9

Effective and timely decisions – Data, information and knowledge – Role of mathematical models – Business intelligence architectures: Cycle of a business intelligence analysis – Enabling factors in business intelligence projects – Development of a business intelligence system – Ethics and business intelligence.

UNIT II KNOWLEDGE DELIVERY 9

The business intelligence user types–Standard reports– Interactive Analysis and Ad Hoc Querying – Parameterized Reports and Self-Service Reporting – dimensional analysis, Alerts/Notifications - Visualization: Charts – Graphs, Widgets – Scorecards and Dashboards – Geographic Visualization – Integrated Analytics – Considerations: Optimizing the Presentation for the Right Message.

UNIT III EFFICIENCY 9

Efficiency measures – The CCR model: Definition of target objectives – Peer groups – Identification of good operating practices – Cross efficiency of analysis – Virtual inputs and outputs – Other models – Pattern matching – Cluster analysis – outlier analysis.

UNIT IV ARCHITECTING THE DATA 9

Introduction, Types of Data, Enterprise Data Model, Enterprise Subject Area Model, Enterprise Conceptual Model, Enterprise Conceptual Entity Model, Granularity of the Data, Data Reporting and Query Tools, Data Partitioning, Metadata, Total Data Quality Management (TDQM).

UNIT V DATA EXTRACTION 9 Introduction, Data Extraction, Role of ETL process, Importance of source identification, Various data extraction techniques, Logical extraction methods, Physical extraction methods, Change data capture.

TOTAL HOURS: 45

TEXT BOOK:

1. Efraim Turban, Ramesh Sharda, DursunDelen, “Decision Support and Business Intelligence Systems”, 9th Edition, Pearson 2013.

REFERENCE(S):

1. Larissa T. Moss, S. Atre, “Business Intelligence Roadmap: The Complete Project Lifecycle of

Decision Making", Addison Wesley, 2003.

2. Carlo Vercellis, "Business Intelligence: Data Mining and Optimization for Decision Making", Wiley Publications, 2009.
3. David Loshin Morgan, Kaufman, "Business Intelligence: The Savvy Manager"s Guide", Second Edition, 2012.
4. CindiHowson, "Successful Business Intelligence: Secrets to Making BI a Killer App", McGraw-Hill, 2007.
5. Ralph Kimball , Margy Ross , Warren Thornthwaite, Joy Mundy, Bob Becker, "The Data Warehouse Lifecycle Toolkit", Wiley Publication Inc.,2007.

COURSE OUTCOMES:

At the end of the course the students will be able to

- CO1 - Explain the fundamentals of business intelligence.
- CO2 - Link data mining with business intelligence.
- CO3 - Apply various modeling techniques.
- CO4 - Explain the data analysis and knowledge delivery stages.
- CO5 - Apply business intelligence methods to various situations.
- CO6 - Decide on appropriate technique.

PO vs CO MAPPING

CO. No	PO_a	PO_b	PO_c	PO_d	PO_e	PO_f	PO_g	PO_h	PO_i	PO_j	PO_k	PO_l
CO1	3	2	2						2			
CO2	3		2		3	3	3					3
CO3	3	2	2	3			3					
CO4	3		2	2								
CO5	3	2	2	3	2	2	2				2	2
CO6												

1→Low 2→Medium 3→High

-o0o-o0o-o0o-