

# Kadi Sarva Vishwavidyalaya, Gandhinagar

## MCA Semester II

### MCA-25 (B) : Big Data and Data Analytics

#### Rationale:

The course provides a deep dive into Big Data Analytics, by giving the fundamental knowledge of the concepts of big data and provides an advanced practical based learning that allows students to lead and develop in Big Data Analytical projects

**Prerequisites:** Knowledge of Database Management Systems, Object Oriented Programming & Basic statistics

#### Learning Outcomes:

- This course will teach how to program in R and use R for effective data analysis.
- The students will learn how to install and configure R necessary for an analytics programming environment and gain basic analytic skills via this high-level analytical language.
- The course covers fundamental knowledge in R programming.
- Students can able to visualize the output in different graphical format
- Popular R packages for data science will be introduced as working examples.

**Teaching and Evaluation Scheme:** The objective of evaluation is to evaluate the students throughout the semester for better performance. Students are evaluated on the basis of continuous evaluation system both in theory and practical classes based on various parameters like term work, class participation, practical and theory assignments, presentation, class test, Regular Attendance, etc.

SUB Total CREDIT	<u>Teaching scheme</u>		<u>Examination scheme</u>				
	(per week)		MID	CEC	External		Total Marks
	Th.	Pr.	Th.	Th.	Th.	Pr.	
5	3	4	25	25	50	50	150

#### Course Contents:

##### Unit 1: The Fundamentals of Big Data

**[20%]**

**Application:** To understand big data concepts, big data adoption, planning and business intelligence

**Understanding Big Data:** Concepts and Terminology, Big data characteristics, Different Types of Data

**Business Motivations and Drivers for Big Data Adoption:** Marketplace Dynamics, Business Architecture, Business Process Management, Information and Communication Technology, Internet of Everything (IoE)

**Big Data Adoption and Planning Considerations:** Organization Prerequisites, Data Procurement, Privacy, Security, Provenance, Limited Realtime Support, Distinct Performance Challenges, Distinct Governance Requirements, Distinct Methodology, Clouds, Big Data Analytics Lifecycle

**Enterprise Technologies and Big Data Business Intelligence:** OLTP, OLAP, ETL, Data Warehouse, Data Marts, Traditional BI, Big Data BI

**Page No: Book – 1: 3 to 20, 29 to 42, 47 to 69, 77 to 87**

## **Unit 2: Introduction and Features of R Language**

**[20%]**

**Application:** To start installing and learn to work with R and RStudio

**What is R?** Installing R, Choosing an IDE, Your First Program, How to Get Help in R, Installing Extra Related Software

**Scientific Calculator:** Mathematical Operations and Vectors, Assigning Variables, Special Numbers, Logical Vectors

**Inspecting Variables:** Classes, Different Types of Numbers, Other Common Classes, Checking and Changing Classes, Examining Variables, Workspace

**Vectors, Matrices and Arrays:** Vectors, Matrices and Arrays

**Lists and Data Frames:** Lists, NULL, Pairlists, Data Frames

**Page No: Book – 2: 2 to 77**

## **Unit 3: Control Flow, Looping, Package, Data Time**

**[20%]**

**Application:** To implement the control flow and looping structure involved in R, to understand and implement different packages and work with date and time

**Environments and Functions:** Environments, Functions

**Strings and Factors:** Strings, Factors

**Flow Control and Loop:** Flow controls, Loops, **Advanced Looping:** Replication, Looping over Lists, Looping Over Arrays, Multiple-Input Apply, Split-Apply-Combine, The plyr Package

**Packages:** Loading Packages, Installing Packages, Maintaining Packages

**Page No: Book – 2: 79 to 150**

## **Unit 4: Working with Date & Time and Data Analysis Workflow**

**[20%]**

**Application:** Working with date time, different set of data and applying cleaning and transformation of data

**Dates and Times:** Date and Time Classes, Conversion to and from Strings, Time Zones, Arithmetic with Dates and Times, Lubridate

**Getting Data:** Built-in Datasets, Reading Text Files, Reading Binary Files, Web Data, Accessing Databases

**Cleaning and Transforming:** Cleaning Strings, Manipulating Data Frames, Sorting, Functional Programming

**Page No: Book – 2: 153 to 202**

## Unit 5: Graphics, Model Creation and Comparison

[20%]

**Application:** To explore and visualize the derived output in different graph format and understanding the graphs, implementation of different distribution and modeling using programming structure

**Exploring and Visualizing:** Summary Statistics, the Three Plotting Systems, Scatterplots, Line Plots, Histograms, Box Plots, Bar Charts, Other Plotting Packages and Systems

**Distributions and Modeling:** Random Numbers, Distributions, Formulae, First Model: Linear Regressions, Other Model Types

**Programming:** Messages, Warnings and Errors, Error Handling, Debugging, Testing

**Page No: Book – 2: 207 to 298**

### Text Book

1. Big Data Fundamentals Concepts, Drivers & Techniques, Thomas Erl, Wajid Khattak, and Paul Buhler, Prentice Hall, Pearson publication
2. Learning R, Richard Cotton, O'Reilly Publications

### Reference Books:

- A Learning Guide to R Beginner to intermediate skills in data analysis, visualization, and manipulation, Remko Duursma, Jeff Powell & Glenn Stone
- R Programming for Data Science, Roger D. Peng, Lean Publishing
- R for Beginners, Emmanuel Paradis
- [http://cran.r-project.org/doc/contrib/Paradis-rdebuts\\_en.pdf](http://cran.r-project.org/doc/contrib/Paradis-rdebuts_en.pdf)
- The Art of R Programming, Norman Matloff
- Big Data Analytics with R and Hadoop, Vignesh Prajapati, PACKT Publication

### Practical Questions:

1. Create two excel file which store the details of the employees personal details, official details, based on the joining details of the employee and the increment given predict list of employee whether promotion can be given or not.
2. Create an excel file to store the details of the patients health details,
  - a. Predict whether the user is health or not based on the health reports.
  - b. Prediction should be done based on the normal values [i.e: BMI, weight, BP, Cholestrol etc]
3. Read two matrixes and perform all the matrix operations like addition, subtraction, multiplication, division, transpose. Assign name for the rows and columns
4. Create multiple vector, read all the vectors, store in a data frame and perform all the operations and conditions based on the vector.
  - a. Assign new name to the data frame
  - b. Attach the new column
  - c. Print the output in the new column based on some conditions
5. Create an excel file which contains the sale details of 3 years in a particular industry.
  - a. Represent the details in the form of histogram, barplot, boxplot

6. Create an excel file which stores the result details of the students of MCA. Predict the next year result based on the criteria like [Theory assignment, Practical assignment, Class performance, attendance, etc ]. Plot the scatter plot of the performance of the students.
7. Write the R code which store the player information like Name, Team, No of times has played, No of goals scored till date.
  - a. Store the details in the .csv file
  - b. Display the details of a single player by entering the name
  - c. Display the full details of a player who has secured maximum score.
  - d. Display the average score of each team.
  - e. Update the score of a particular team and store the details in .csv through R
8. Perform the list of operations for the following:
  - a. List the objects in memory.
  - b. Clear the screen.
  - c. Declare variables x, y and assign values of 5 and 8 to x and y.
  - d. Perform simple calculations like addition, subtraction, division, multiplication etc. on x and y.
  - e. Print the values of variables on screen.
  - f. Assign five distinct values to z.
  - g. Assign sequential value from 1 to 20
  - h. Declare an array a.
  - i. Input multiple values from the user at prompt and store it in c.
  - j. Show the data types of all objects on screen.
  - k. Sort the values in descending order.
  - l. Find out the sum, max, min, diagonal element of matrix.
  - m. Find out the working directory and change it.
  - n. Remove x and y objects from memory.
  - o. Print only odd numbers of series.
9. Create a matrix of 3 x 3 and make layout, and print the data in the layout.
10. Generate a graphical image by using all plot, define the title, x-axis, y-axis, x limit and y-limit of a graph for a .csv file?
11. Retrieve the data from the .csv file
  - a. Normalize the data
  - b. Represent in a graphical form
  - c. Specify x-axis, y-axis, x-limit, y-limit, include color to the graph, change the plot style
12. Use the lattice library and display the graphical image of all lattice form
13. Generate the .csv file, create different models and specify the,
  - a. Summary of different models
  - b. Find the residual, co-efficient, fitted and AIC.
14. Write a R function to calculate the Fibonacci series.
15. Write the R code to predict whether loan to be sanctioned or not to be sanctioned for a particular customer.
  - a. Prepare the dataset of list of customers with their personal data, salary details, previous loan taken, EMI details per month, bank account details.
  - b. Based on the input criteria predict whether loan to be given or not.
  - c. If sanctioned mention the loan amount that got sanctioned
  - d. Prepare a separate file and store the output details
  - e. Display the current years currents loan status in a graph