I YEAR II SEMESTER PAPER– II

INTRODUCTION TO DATA SCIENCE WITH R

Objective

Data Science is a fast-growing interdisciplinary field, focusing on the analysis of data to extract knowledge and insight. This course will introduce students to the collection. Preparation, analysis, modeling and visualization of data, covering both conceptual and practical issues. Examples and case studies from diverse fields will be presented, and hands-on use of statistical and data manipulation software will be included.

Outcomes

1. Recognize various disciplines that contribute to a successful data science effort.
2. Understand the processes of data science - identifying the problem to be solved, data collection, preparation, modeling, evaluation and visualization.
3. Be aware of the challenges that arise in data sciences.
4. Develop and appreciate various techniques for data modeling and mining.
5. Be cognizant of ethical issues in many data science tasks.
6. Be comfortable using commercial and open source tools such as the R language and its associated libraries for data analytics and visualization.
7. Learn skills to analyze real time problems using R
8. Able to use basic R data structures in loading, cleaning the data and preprocessing the data.
9. Able to do the exploratory data analysis on real time datasets
10. Able to understand and implement Linear Regression
11. Able to understand and use - lists, vectors, matrices, dataframes, etc.

Unit-1:

Introduction to Data Science- Introduction- Definition - Data Science in various fields - Examples - Impact of Data Science - Data Analytics Life Cycle - Data Science Toolkit - Data Scientist - Data Science Team

Understanding data: Introduction – Types of Data: Numeric – Categorical – Graphical – High Dimensional Data – Classification of digital Data: Structured, Semi-Structured and Un-Structured - Example Applications. Sources of Data: Time Series – Transactional Data – Biological Data – Spatial Data – Social Network Data – Data Evolution.

Unit-2:

Introduction to R- Features of R - Environment - R Studio. Basics of R-Assignment - Modes - Operators - special numbers - Logical values - Basic Functions - R help functions - R Data Structures - Control Structures. Vectors: Definition- Declaration - Generating - Indexing - Naming - Adding & Removing elements - Operations on Vectors - Recycling - Special Operators - Vectorized if- then else-Vector Equality – Functions for vectors - Missing values - NULL values - Filtering & Subsetting.

Unit-3:

Matrices - Creating Matrices - Adding or Removing rows/columns - Reshaping - Operations - Special functions on Matrices. Lists - Creating List – General List Operations - Special Functions - Recursive Lists. Data Frames - Creating Data Frames - Naming - Accessing -

Adding - Removing - Applying Special functions to Data Frames - Merging Data Frames- Factors and Tables.

Unit- 4:

Input / Output – Reading and Writing datasets in various formats - Functions - Creating User-defined functions - Functions on Function Object - Scope of Variables - Accessing Global, Environment - Closures - Recursion. Exploratory Data Analysis - Data Preprocessing - Descriptive Statistics - Central Tendency - Variability - Mean - Median - Range - Variance - Summary - Handling Missing values and Outliers - Normalization
Data Visualization in R : Types of visualizations - packages for visualizations - Basic Visualizations, Advanced Visualizations and Creating 3D plots.

Unit- 5:

Inferential Statistics with R - Types of Learning - Linear Regression- Simple Linear Regression - Implementation in R - functions on lm() - predict() - plotting and fitting regression line. Multiple Linear Regression - Introduction -comparison with simple linear regression - Correlation Matrix - F-Statistic - Target variables Vs Predictors - Identification of significant features - Implementation of Multiple Linear Regression in R.

References

1.Nina Zumel, John Mount, "Practical Data Science with R", Manning Publications, 2014. 2.Jure Leskovec, Anand Rajaraman, Jeffrey D.Ullman, "Mining of Massive Datasets",
Cambridge University Press, 2014.
3.Mark Gardener, "Beginning R - The Statistical Programming Language", John Wiley & Sons, Inc., 2012.
4.W. N. Venables, D. M. Smith and the R Core Team, "An Introduction to R", 2013. 5.Tony Ojeda, Sean Patrick Murphy, Benjamin Bengfort, Abhijit Dasgupta, "Practical Data
Science Cookbook", Packt Publishing Ltd., 2014.
6. Nathan Yau, "Visualize This: The FlowingData Guide to Design, Visualization, and Statistics", Wiley, 2011.
7. Boris lublinsky, Kevin t. Smith, Alexey Yakubovich, "Professional Hadoop Solutions", Wiley, ISBN: 9788126551071, 2015.

Student Activity

Databases need to undergo pre-processing to be useful for data mining. Dirty data can cause confusion for the data mining procedure, resulting in unreliable output. Data cleaning includes smoothing noisy data, filling in missing values, identifying and removing outliers, and resolving inconsistencies.

RECOMMENDED CO-CURRICULAR ACTIVITIES:

(Co-curricular activities shall not promote copying from textbook or from others work and shall encourage self/independent and group learning)

A. Measurable

1. Assignments (in writing and doing forms on the aspects of syllabus content and outside the syllabus content. Shall be individual and challenging)

2. Student seminars (on topics of the syllabus and related aspects (individual activity))

3. Quiz (on topics where the content can be compiled by smaller aspects and data (Individuals or groups as teams))

4. Study projects (by very small groups of students on selected local real-time problems pertaining to syllabus or related areas. The individual participation and contribution of students shall be ensured (team activity

B. General

1. Group Discussion

2. Try to solve MCQ's available online.

3. Others

RECOMMENDED CONTINUOUS ASSESSMENT METHODS:

Some of the following suggested assessment methodologies could be adopted;

1. The oral and written examinations (Scheduled and surprise tests)

2. Closed-book and open-book tests

3. Problem-solving exercises

4. Practical assignments and laboratory reports

5. Observation of practical skills

6. Individual and group project reports like "COVID-19 Analysis", "Estimated Quanrantain Period for Covid-19 Contacts", etc.

7. Efficient delivery using seminar presentations,

8. Viva voce interviews.

9. Computerized adaptive testing, literature surveys and evaluations,

10. Peers and self-assessment, outputs form individual and collaborative work

I YEAR II SEMESTER PAPER– II

R Programming LAB

1) Installing R and R studio
2) Create a folder DS_R and make it a working directory. Display the current working directory
3) installing the "ggplot2", "caTools", "CART" packages

4) load the packages "ggplot2", "caTools".
5) Basic operations in r
6) Working with Vectors:
   - Create a vector v1 with elements 1 to 20.
   - Add 2 to every element of the vector v1.
   - Divide every element in v1 by 5
   - Create a vector v2 with elements from 21 to 30. Now add v1 to v2.
7) Getting data into R, Basic data manipulation
8) Using the data present in the table given below, create a Matrix "**M**"

	C1	*C2*	*C3*	*C4*	*C5*
C1	*0*	*12*	*13*	*8*	*20*
C2	*12*	*0*	*15*	*28*	*88*
C3	*13*	*15*	*0*	*6*	*9*
C4	*8*	*28*	*6*	*0*	*33*
C5	*20*	*88*	*9*	*33*	*0*

   - Find the pairs of cities with shortest distance.
9) Consider the following marks scored by the 6 students

Section	Student no	M1	M2	M3
A	1	45	54	45
A	2	34	55	55
A	3	56	66	64
B	1	43	44	45
B	2	67	76	78
B	3	76	68	37

- create a data structure for the above data and store in proper positions with proper names
- display the marks and totals for all students
- Display the highest total marks in each section.
- Add a new subject and fill it with marks for 2 sections.

- Three people denoted by P1, P2, P3 intend to buy some rolls, buns, cakes and bread. Each of them needs these commodities in differing amounts and can buy them in two shops S1, S2. The individual prices and desired quantities of the commodities are given in the following table "demand.

	price	
	S1	S2
Roll	1.5	1
Bun	2	2.5
Cake	5	4.5
Bread	16	17

	demand.quantity			
	Roll	**Bun**	**Cake**	**Bread**
P1	6	5	3	1
P2	3	6	2	2
P3	3	4	3	1

- Create matrices for above information with row names and col names.
- Display the demand.quantity and price matrices
- Find the total amount to be spent by each person for their requirements in each shop
- Suggest a shop for each person to buy the products which is minimal.

10) Consider the following employee details:

employee details as follows		
	emp_no:1	
	name: Ram	
	salary	
		basic: 10000
		hra: 2500
		da: 4000
	deductions	
		pf: 1100
		tax: 200
	total salary	
		gs(Gross Salary):
		ns(Net Salary)

- Create a list for the employee data and fill gross and net salary.
- Add the address to the above list
- display the employee name and address
- remove street from address
- remove address from the List.

11) Loops and functions - Find the factorial of a given number
12) Implementation of Data Frame and its corresponding operators and functions
13) Implementation of Reading data from the files and writing output back to the specified file
14) Treatment of NAs, outliers, Scaling the data, etc
15) Applying summary() to find the mean, median, standard deviation, etc
16) Implementation of Visualizations - Bar, Histogram, Box, Line, scatter plot, etc.
17) Implementation of Linear and multiple Linear Regression
18) Fitting regression line