



**SARVAJANIK UNIVERSITY**  
**SarvajaniK College of Engineering and Technology**  
**Bachelor of Technology**  
**Instrumentation and Control**



**B. Tech. III Semester –VI**  
**Program: Honors- Industry 4.0**

**Subject Name: Fundamental of Data Science**  
**Type of course: Honors**

**Subject Code: BTIC19621**

**Prerequisite (if any):**

**List of Courses where this course will be prerequisite:** Data Analysis and Visualisation,DCS

**Rationale:**

This course will provide the students platform for learning tools which helps in data gathering, processing and analysis . It also includes some basic data handling algorithm which helps in prediction.

**Teaching and Examination Scheme:**

TEACHING SCHEME				Theory Marks			Practical Marks		Total
L	T	P	C	TEE	CA1	CA2	TEP	CA3	
3	0	2	4	60	25	15	30	20	150

**CA1: Continuous Assessment (assignments/projects/open book tests/closed book tests** CA2: Sincerity in attending classes/class tests/ timely submissions of assignments/self-learning attitude/solving advanced problems TEE: Term End Examination TEP: Term End Practical Exam (Performance and viva on practical skills learned in course) CA3: Regular submission of Lab work/Quality of work submitted/Active participation in lab sessions/viva on practical skills learned in course

Sr. No.	Content	Total Hrs	Module Weightage

**BSC: basic science course /ESC: Engineering Science Course /HSM: Humanities and management /PCC: Professional Core course /PEC: professional Elective course /OEC: Open Elective course/ MD: mandatory noncredit course**

<b>1</b>	<b>Introduction to Data Science</b> Component of Data Science, Data Science Process, Data Science Roles, Difference between Data Science and Business Intelligence, Applications of Data Science in various fields, Challenges of Data Science Technology	<b>2</b>	<b>4%</b>
<b>2</b>	<b>Data Collection and Data Pre-Processing</b> Data Collection Methods, Data Pre-Processing Overview, Data Cleaning, Data Integration and Transformation, Data Reduction, Data Discretization	<b>5</b>	<b>11%</b>
<b>3</b>	<b>Model Development</b> Linear Regression (Simple and Multiple Regression) , Logistic Regression, Ridge regression, Lasso Regression, Polynomial Regression, Stepwise Regression	<b>8</b>	<b>18%</b>
<b>4</b>	<b>Model Evaluation</b> Generalization Error, Out-of-Sample Evaluation Metrics, Cross Validation, Over fitting, Under Fitting, Prediction by using Ridge Regression, Testing Multiple Parameters by using Grid Search	<b>6</b>	<b>13%</b>
<b>5</b>	<b>Getting Started with Pandas</b> Arrays and vectorized computation, Introduction to pandas Data Structures, Essential Functionality, Summarizing and Computing Descriptive Statistics.  <b>Data Loading, Storage and File Formats</b> Reading and Writing Data in Text Format, Web Scraping, Binary Data Formats, Interacting with Web APIs, Interacting with Databases  <b>Data Cleaning and Preparation</b> Handling Missing Data, Data Transformation, String Manipulation	<b>10</b>	<b>22%</b>

<b>6</b>	<b>Data Wrangling:</b> Hierarchical Indexing, Combining and Merging Data Sets Reshaping and Pivoting.	<b>4</b>	<b>8%</b>
<b>7</b>	<b>Data Aggregation and Group operations:</b> Group by Mechanics, Data aggregation, General split-apply-combine, Pivot tables and cross tabulation	<b>4</b>	<b>8%</b>
<b>8</b>	<b>Mining Data Stream</b> Sampling Data in a Stream, Filtering Streams, Counting Distinct Elements in a Stream, Clustering for Streams	<b>6</b>	<b>13%</b>

**Content:**

**Suggested Specification table with Marks (Theory): (For B Tech only)**

Distribution of Theory Marks					
R Level	U Level	A Level	N Level	E Level	C Level
<b>10%</b>	<b>10%</b>	<b>10%</b>	<b>20%</b>	<b>10%</b>	<b>40%</b>

**Legends: R: Remembrance; U: Understanding; A: Application, N: Analyze and E: Evaluate C: Create and above Levels (Revised Bloom’s Taxonomy)**

Note: This specification table shall be treated as a general guideline for students and teachers. The actual distribution of marks in the question paper may vary slightly from above table.

**Reference Books:**

Sr no	Title of book /article	Author(s)	Publisher and details like ISBN	Year of publication	Publication Edition
1	Data science from scratch	Joel Grus	O'Reilly Media, ISBN:	2015	1st Edition

**BSC: basic science course /ESC: Engineering Science Course /HSM: Humanities and management /PCC: Professional Core course /PEC: professional Elective course /OEC: Open Elective course/ MD: mandatory noncredit course**

			9781491901410		
2	Practical Statistics for Data Scientists: 50	Peter Bruce, Andrew Bruce	O'reilly publishing house ISBN: 9781491952962	2017	1st Edition
3	Doing Data Science	Cathy O'Neil, Rachel Schutt	O'Reilly Media, ISBN: 9781449358655	2013	1st Edition

**Course Outcomes:**

After learning this course, students will be able to

Sr. No.	CO statement	Marks weightage %
CO-1	Perform descriptive statistics and dimensionality reduction.	10%
CO-2	Discuss and apply various data science approaches.	15%
CO-3	Apply various data pre-processing and manipulation techniques	25%
CO-4	Develop the model and apply methods for model evaluation	25%
CO-5	Interpret and evaluate various large datasets by applying data aggregation, Data Mining techniques like clustering, filtering, factorization.	25%

How strongly Pos are addressed by each CO of a subject is indicated by level 1: Slight (Low) 2: Moderate (Medium) 3: Substantial (High)

**Mapping CO-POs**

	PO 1	PO 2	PO 3	PO4	PO 5	PO 6	PO 7	PO8	PO9	PO10	PO11	PO12	PSO1	PSO2	PSO3
CO1	1			3	3							3	3		2

**BSC: basic science course /ESC: Engineering Science Course /HSM: Humanities and management /PCC: Professional Core course /PEC: professional Elective course /OEC: Open Elective course/ MD: mandatory noncredit course**

<b>CO2</b>	1			3	3							3	3		2
<b>CO3</b>	1			3	3							3	3		2
<b>CO4</b>	1			3	3					2		3	3		3
<b>CO5</b>	2			3	3					3		3	3		3

**List of Open learning website:**

[www.nptel.ac.in](http://www.nptel.ac.in)

**FOR LAB SESSIONS:**

**List of Experiments:**

Data Analysis practicals will be based on discussion in the classroom.

1. Perform experiments using NumPy.
2. Explore Pandas Data Structures.
3. Write a program for Data Loading, Storage and File Formats.
4. Perform experiments based on Data Cleaning and Preparation.
5. Explore Data Wrangling.
6. Explore data mining.
7. For a given database, model using linear regression and evaluate accuracy of them.
8. For a given database, model using logistic regression and evaluate accuracy of them.

**Major Equipment Needed:** Computers