SARVAJANIK UNIVERSITY
INCLUSIVE | INTEGRATED | INNOVATIVE

SARVAJANIK UNIVERSITY
Sarvajanik College of Engineering and Technology
Bachelor of Technology

Year: B. Tech III (Semester V)

Subject Name: Fundamentals of Data Science **Subject Code:** BTAI13502
Type of course: Professional Core Course
Prerequisite (if any): Probability and Statistics

Rationale: This course provides insights into the data analysis and fundamentals of Data Science. Students can gain a thorough understanding of data science life cycle, data analytics and tools. Students will be able to solve predictive problems on real time data after learning this course.

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.

Contents:

Sr. No.	Contents	Total Hrs
1.	**Introduction to Data Science** : Fundamentals of data science, life cycle, data types, tools and techniques, applications.	03
2.	**Exploratory Data Analysis:** Elements of Structured Data, Data Frames and Indexes, Estimates of Location, Estimates of Variability, Exploring the Data Distribution, Exploring Binary and Categorical Data, Correlation, Exploring Two or More Variables.	05
3.	**Data and Sampling Distributions:** Random sampling and sample Bias, Selection Bias, Sampling Distribution of a Statistic, The Bootstrap, Confidence Intervals, Normal Distribution, Students' t-Distribution, Binomial Distribution, Chi-Square Distribution, F-Distribution.	08
4.	**Statistical Inference:** Hypothesis Tests, Resampling, Statistical Significance and p-Values, t-Tests, Degrees of Freedom, ANOVA, Chi-Square Test.	08

5.	**Model Building and Evaluation:** Simple Linear Regression, Residual sum of squares (RSS), Least Square, Multiple Linear Regression, Prediction Using Regression, Interpreting the Regression Equation, Regression Diagnostics, Logistic Regression -Logistic Response Function and Logit, Generalized Linear Models, Predicted Values from Logistic Regression, Interpreting the Coefficients and Odds Ratios, Similarities and Differences between Linear and Logistic Regression, Assessing the Model, Evaluating Classification Models - Confusion Matrix, The Rare Class Problem, Precision, Recall, and Specificity, ROC Curve, AUC, Lift .	**15**
6.	**Introduction to R :** Variables and data types in R, Data frames, Recasting and joining of data frames, Arithmetic, Logical and Matrix operations in R, Advanced programming in R: Functions, Control structures, Data visualization in R Basic graphics.	**06**

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**
15	**25**	**10**	**10**	**-**	**-**

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

Reference Books:

Sr no	**Title of book /article**	**Author(s)**	**Publisher and details like ISBN**
1.	Practical Statistics for Data Scientists	Peter Bruce and Andrew Bruce	O'Reilly Media, Inc.
2.	Think Stats: Probability and Statistics for Programmers	Allen B. Downey	Green Tea Press, Needham, Massachusetts
3.	Data Science for Dummies	John Muller	Wiley
4.	Statistics for Data Science	James D. Miller	PACKT

Course Outcomes (CO):

Sr. No.	CO statements	Marks % weightage
CO-1	Understand categories of data, Analyse and summarize them with the help of various exploratory data analysis approaches	20%
CO-2	Understand various sampling distributions, apply suitable test to identify statistically significant relationship between categorical variables	30%
CO-3	Build models for data analytics and apply different evaluation metrics to understand the model performance.	40%
CO-4	Understand and visualize data with graphical techniques like barplot, histogram, using R	10%

List of Open learning website:

- https://mml-book.com/
- https://www.math.ucdavis.edu/~linear/
- https://ibse.iitm.ac.in/course/math-foundations-of-ds/
- NPTEL Course on Python for Data Science : https://onlinecourses.nptel.ac.in/noc21_cs33/
- Course on Data Science with Python : https://www.simplilearn.com/getting-started-data-science-with-python-skillup

List of Suggested Experiments:

Consider dataset with student name, gender, Enrolment no, semester result with marks of each subject, his mobile number, city. Implement following in Python or R.

1. Perform descriptive analysis and identify the data type.
2. Implement a method to find out variation in data. For example the difference between highest and lowest marks in each subject semester wise.
3. Plot the graph showing result of student in each semester.
4. Plot the graph showing the geographical location of students.
5. Plot the graph showing number of male and female students.
6. Implement a method to treat missing value for gender and missing value for marks.
7. Implement linear regression to predict the 5th semester result of student.
8. Implement logistic regression to classify the student as average or clever.