SARVAJANIK UNIVERSITY
INCLUSIVE | INTEGRATED | INNOVATIVE

SARVAJANIK UNIVERSITY
Sarvajanik College of Engineering and Technology
Bachelor of Technology

Year: B. Tech II (Semester IV)

Subject Name: Statistics and Data Preprocessing **Subject Code:** BTEA19421
Type of course: Honors (Group: Data Science)
Prerequisite (if any): Python Programming, Probability and Statistics

Rationale: This Statistics and Data Preprocessing course is designed to introduce students to the basic principles of statistical methods and techniques used for data preprocessing. After completing this course students will have theoretical and practical knowledge of topics in statistics including - data gathering, data summarization, examining relationships between variables, introduction to ANOVA (analysis of variance), regression and correlation analysis. Students will take a hands-on approach to statistical analysis using Python and Jupyter Notebooks – the tools of choice for Data Scientists and Data Analysts.

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

Content:

Sr. No.	Contents	Total Hours
1	**Introduction to Data Science** : Fundamentals of data science, life cycle, data types, tools and techniques, applications	03
2	**Descriptive Statistics:** Introduction to probability, Probability Distribution Function, Density Distribution Function, Sampling and sampling distributions for discrete and continuous random variable, central limit theorem, random number generator	07
3	**Categorical and Quantitative Description:** Count, Mean, Median, Mode, Variance, Standard Deviation, skewness, kurtosis, range, inter quartile range, five number summary, outlier detection	07

4	**Analysis of variance (ANOVA)** : Pearson Correlation, Covariance , Chi Square Test, One and Two Tail Test, F - Score Calculations, MAE, MSE, RMSE, PSNR	08
5	**Data Pre-processing:** Determining instances and Features, identifying data types, rescaling, standardization, normalization - z- score, min-max, decimal scaling, one-hot encoding, binary encoding, label encoding	10
6	**Exploratory Data Analysis (EDA):** Univariate/Bivariate/Multivariate Analysis, removing irrelevant features, finding duplicate datas, missing value analysis, outlier treatment, variable transformation, variable creation	10

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	20	0	0	0

Legends: R: Remembrance; U: Understanding; A: Application, N: Analyze and E: Evaluate C: Create (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.	Think Stats: Probability and Statistics for Programmers	Allen B. Downey	Green Tea Press Needham, Massachusetts	2014	2nd Edition
2.	Data Science for Dummies	John Muller	Wiley	2017	2nd Edition
3.	Practical Statistics for Data Scientists	Peter Bruce and Andrew Bruce	Publisher(s): O'Reilly Media, Inc. ISBN: 9781491952962	2017	2nd Edition
4.	Statistics for Data Science	James D. Miller	PACKT	2017	-

Course Outcomes:

Sr. No.	CO statements	Marks % weightage
CO-1	Analyze the probability distribution function for discrete and continuous random variables	25
CO-2	Describe various descriptive statistics for categorical and quantitative data	15
CO-3	Demonstrate Analysis of Variance (ANOVA) statistical method to find the correlation among two or more groups of data.	15
CO-4	Apply various data preprocessing steps in order to improve quality of data.	20
CO-5	Apply exploratory data analysis to summarize the main characteristics of the data sets.	25

List of Open learning website:

1. Data Analytics with Python : https://onlinecourses.nptel.ac.in/noc21_cs45/course
2. NPTEL Course on Python for Data Science : https://onlinecourses.nptel.ac.in/noc21_cs33/
3. Coursera Specialization Course on Applied Data Science with Python Programming : https://www.coursera.org/specializations/data-science-python

List of Open Source Software:

1. Scikit-Learn : Stats, pdf
2. Pandas

For Lab Sessions:

List of Experiments:

Sr. No	Practical
1.	Consider the given data and find the five summary point for it.
2.	Generate random data and find the probability distribution for the same
3.	Write a program to calculate mean, median, mode, standard deviation for given datasets.
4.	Write a program to detect outliers using IQR from randomly generated sufficient size data (Minimum size of data should be 100).

5.	With the mentioned values of x and y, create the following models, plot them and find out their score value. Compare and contrast the two tail tests with respect to F -score. x= np.array([[-3,7],[1,5], [1,2], [-2,0], [2,3], [-4,0], [-1,1], [1,1], [-2,2], [2,7], [-4,1], [-2,7]]) y = np.array([3, 3, 3, 3, 4, 3, 3, 4, 3, 4, 4, 4])
6.	Generate Binomial PDF for the random variates with n=20, p=0.7 and size=1000.
7.	Generate Poisson PDF for the random variates with λ=20, and size=10000.
8.	Create an array of 1 million random numbers in a specific range (say -5 to 5) from uniform distribution. Create an array of random numbers from normal distribution. Generate Normal & Uniform distribution for the generated numbers.
9.	Mini project to implement the data preprocessing and EDA.