# SARVAJANIK UNIVERSITY
# Sarvajanik College of Engineering and Technology
# Bachelor of Technology

## Year: B. Tech III (Semester V)

**Subject Name:** Data Science using Python **Subject Code:** BTIT15501
**Type of course:** Open Elective Course
**Prerequisite (if any):** Programming for Problem Solving

**Rationale:** This course covers use of python programming for solving data science problems.

**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 Hours |
|---|---|---|
| 1. | **Introduction to data analytics :**Data analytics vs. Data analysis**,** Types of Data analytics- Descriptive analytics, Diagnostic analytics, Predictive analytics, Prescriptive analytics, Categories of Variables: Nominal, Ordinal, Interval, Ratio | 04 |
| 2. | **Getting Started with Raw Data:**<br>**The world of arrays with NumPy:** Creating an array, Mathematical operations, Array subtraction, Squaring an array, A trigonometric function performed on the array, Conditional operations, Matrix multiplication, Indexing and slicing, Shape manipulation,<br>**Empowering data analysis with pandas**: the data structure of pandas, Series, Inserting and exporting data-CSV, XLS, **Data cleansing:** Checking the missing data, Filling the missing data, String operations, Merging data, **Data operations:** Aggregation operations | 07 |
| 3. | **Plotting and Visualization**<br>A Brief matplotlib API Primer, Figures and Subplots, Colors, Markers, and Line Styles, Ticks, Labels, and Legends, Annotations and Drawing on a Subplot, Saving Plots to File, matplotlib Configuration, Plotting Functions in pandas: Line Plots, Bar Plots, Histograms and Density Plots, Scatter Plots | 05 |
| 4. | **Inferential Statistics:**<br>**Measures of Central Tendency:** Arithmetic mean, Weighted mean, Median, Percentile, Dispersion, Skewness, Kurtosis, Range, Interquartile range, Variance, Coefficient of variation, **Probability Distributions:** Expected Value, Variance and Standard Deviation, | 08 |

| | | |
|---|---|---|
| | Covariance, Correlation Coefficient, The Normal Distribution properties, **Hypothesis Testing and ANOVA:** Null and Alternative Hypotheses, Type I and Type II Error, Analysis of Variance(ANOVA) | |
| 5. | **Linear Regression**: Simple Linear Regression Model, Least Squares Method, Coefficient of Determination, Testing for Significance, Using the Estimated Regression Equation for Estimation and Prediction, Sum of squares and sum of cross-products, Coefficient of Determination, Correlation Coefficient, Multiple Linear regression, Adjusted Multiple Coefficient of Determination, F test significance, t Test for individual significance, **Logistic Regression:** Objective of logistic regression, Logistic Regression equation, Applications, Testing for Significance, G Statistics, Linear Regression Model Vs Logistic Regression Model, **Confusion matrix and ROC**: Accuracy, Recall, Precision, F-Score, Specificity, sensitivity, Area Under the ROC Curve (AUC) | 10 |
| 6. | **Cluster Analysis :** Cluster and discriminant analysis, Standardizing the data, Detecting outlier, Distances computation between the objects: Euclidean distance, Manhattan distance, Minkowski distance, Interpretation of distance matrix, Similarities and dissimilarities, Handling different types of variables, K-Means Clustering, Hierarchical Agglomerative Clustering (HAC) | 06 |
| 7. | **Classification and Regression Trees:** Decision Tree Algorithm, splitting criterion, termination conditions, Attribute Selection Measures: information gain, gain ratio, and Gini index | 05 |

**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 | 20 | 20 | 5 | - | - |

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 | Mastering Python for Data Science | Samir Madhavan | Ingram short title |
| 2 | Python for data analysis: Data wrangling with Pandas, NumPy, and IPython. | McKinney, W. | O'Reilly Media, Inc. |
| 3 | Applied Statistics & Probability for Engineering. | Douglas Montgomery, George Runger | John Wiley & Sons, Inc |
| 4 | Practical Statistics for Data Scientists | Peter Bruce and Andrew Bruce | O'Reilly Media |
| 5 | Data Science for Dummies | John Muller | Wiley |

**Note: Students should refer to the latest editions of books**

**Course Outcomes (CO):**

| Sr. No. | CO statements | Marks % weightage |
|---|---|---|
| CO-1 | Understand basics of Python its data handling and visualization operations. | 25% |
| CO-2 | Understand the inferential statistics and data distribution characteristics | 20% |
| CO-3 | Able to develop various regression models and analyse its performance measures | 20% |
| CO-4 | Understand and build clustering and classification models. | 25% |

**List of Online Learning Resources:**

https://onlinecourses.nptel.ac.in/noc23_cs08/course

**List of Experiments:**

Students have to perform minimum 10 laboratory practical based on contents of the course