PRASAD V. POTLURI SIDDHARTHA INSTITUTE OF TECHNOLOGY

KANURU, VIJAYAWADA

III B.Tech – I Sem

DATA SCIENCE

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Offering Branches | CSE | Course Code: | | 20CS4501A |
| Course Category: | Program Elective Course | Credits: | | 3 |
| Course Type: | Theory | Lecture-Tutorial- Practical: | | 3-0-0 |
| Prerequisites: | Engineering Mathematics -2  (Probability & Statistics) | Continuous Evaluation: | | 30 |
| Semester  End Evaluation: | | 70 |
| Total Marks: | | 100 |
| **COURSE OUTCOMES** | | | | |
| Upon successful completion of the course, the student will be able to: | | | | |
| **CO1** | Understand the life cycle process of data science. | | **L2** | |
| **CO2** | Apply different data pre-processing techniques for improving data quality. | | **L3** | |
| **CO3** | Apply statistical methods to evaluate the data. | | **L3** | |
| **CO4** | Apply Statistical Learning techniques for model building, Assesment and Selection. | | **L3** | |
| **Course Contents** | | | | |
| **Unit-1** | **Introduction to Data Science-**  What is Data Science?  Phases of Data Science: Data Acquisition, Cleansing, Exploratory Data Analysis, Data Preparation, Data Modeling.  Engineering Aspects of Data Science: Business Understanding, Data Understanding, Data Preparation, Model Building, Model Evaluation, Hyper Parameter Optimization and Deployment. | | **CO1** | |
| **Unit-2** | **Data Preprocessing:**  Introduction, Data Quality, Data Cleaning- Missing Values, Noisy data, Data Integration, Data Transformation- Smoothing, Attribute construction, Aggregation, Normalization, Discretization, Data Reduction- Wavelet Transforms, Principal Components Analysis, Attribute Subset Selection, Histograms, Clustering, Sampling | | **CO1, CO2** | |
| **UnIt-3** | **Random Variables and Probability Distributions:**  Random variables (discrete and continuous), Probability Density Function (PDF), Probability Mass Function (PMF), and Cumulative Density Function (CDF). Discrete distributions- Uniform, Binomial, Bernoulli and Poisson distributions. Continuous Distributions- Normal distribution, Standard Normal distribution, Student's T distribution, Chi-squared distribution.  Sampling Strategies: Introduction, Simple Random sampling, Systematic sampling, Stratified sampling, Cluster sampling. | | **CO1, CO3** | |
| **Unit-4** | **Linear methods for Regression:**  Introduction, Linear Regression models, Least Squares, Multiple Regression.  Linear methods for Classification: Introduction, Linear discriminative analysis, Logistic Regression. | | **CO1, CO4** | |
| **Unit-5** | **Model Assessment and Selection:**  Introduction, Bias, Variance and Model complexity, Bias-Variance decomposition, Optimism of the training error rate, Estimates of in-sample prediction error, Effective number of parameters, minimum description length, Holdout sets, and cross-validation. | | **CO1, CO4** | |

|  |
| --- |
| **Learning Resources** |
| **Text Books** |
| 1. Introducing Data Science, David Cielen, Arno D. B. Meysman, and Mohamed Ali, 2016, Manning Publications. (UNIT-I) 2. Data Mining: Concepts and Techniques, Jiawei Han, Micheline Kamber and Jian Pei, Third edition, Morgan Kaufmann. (UNIT-II) 3. The Elements of Statistical Learning, Trevor Hastie, Robert Tibshirani, Jerome Friedman, Second Edition, Springer. (UNIT-III, IV, V) |
| **References** |
| 1. Cathy O'Neil and Rachel Schutt, "Doing Data Science", O'Reilly, 2015. 2. Data Science from Scratch: First Principles with Python, Joel Grus, Second edition, 2019, O'Reilly 3. Statistics, Robert S. Witte and John S. Witte, Eleventh Edition, 2017, Wiley Publications. |
| **e- Resources & Other digital material** |
| 1. <https://nptel.ac.in/courses/106106212> 2. <https://nptel.ac.in/courses/106106179> 3. Data Science Methodology- Coursera - <https://www.coursera.org/learn/datascience-methodology> 4. Foundations of Data Science - edX - <https://www.edx.org/course/foundationsof-data-science> |