# 19CS42O1 - DATA ANALYTICS

|  |  |  |  |
| --- | --- | --- | --- |
| Course Category: | Open Elective | Credits: | 3 |
| Course Type: | Theory | Lecture-Tutorial-Practical: | 3-0-0 |
| Prerequisite: | Basic concepts of Data base Management Systems and Knowledge of Probability and Statistics | Sessional Evaluation:  Univ. Exam Evaluation:  Total Marks: | 40  60  100 |
| Objectives: | Students undergoing this course are expected to understand:   * To learn the principles and methods of statistical analysis * Discover interesting patterns, analyze supervised and unsupervised models and estimate the accuracy of the algorithms. * To understand the techniques of time series and text analysis. | | |

|  |  |  |
| --- | --- | --- |
| Course Outcomes | Upon successful completion of the course, the students will be able to: | |
| CO1 | Understand the importance of learning the basics of Big Data analytics |
| CO2 | Learn the fundamentals of R and Hadoop to develop simple data analysis applications |
| CO3 | Learn and understand the various statistical methods |
| CO4 | Understand the basic concepts of supervised and unsupervised models |
| CO5 | Study and understand the time series analysis |
| CO6 | Learn and understand the text analysis and sentiment analysis. |
| Course Content | UNIT-I  **Introduction to Big Data Analytics:**  Big Data Overview, State of the Practice in Analytics, Key Roles for the New Big Data Ecosystem, Examples of Big Data Analytics.  **Data Analytics Life Cycle:**  Data Analytics Lifecycle Overview, Discovery, Data Preparation, Model Planning, Model Building, Communicate Results, Operationalize, Case Study: Global Innovation Network and Analysis (GINA).  UNIT-II  **Analytic Methods Using R:**  Introduction to R, Exploratory Data Analysis, Statistical Methods for Evaluation.  **Advanced Analytics:**  Analytics for Unstructured Data- MapReduce and Hadoop, The Hadoop Ecosystem, SQL essentials.  UNIT-III  Regression:  Categorical Variable, Linear Regression, Logistic Regression, Ordinary Least Squares (OLS), Receiver Operating Characteristic (ROC) Curve, Residuals.  UNIT-IV  **Clustering:**  Overview, K-Means, PAM, Density-Based Clustering  **Classification:**  Decision Tress, Naïve Bayes, Diagnostics of Classifiers  UNIT-V  **Time Series Analysis:**  Overview, ARIMA Model, Building and evaluating ARIMA Model, Additional Methods-ARMAX, GARCH  UNIT-VI  **Text Analysis:**  Text Analysis Steps, Collecting Raw Text, Representing Text, Term Frequency-Inverse Document Frequency (TDIDF), Categorizing Documents by Topics, Determining Sentiments, Gaining Insights. | |
| Text Books &  References  Books | **TEXT BOOKS**   1. Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data by EMC2 Education Services, Published by John Wiley & Sons, Inc.   **REFERENCE BOOKS**   1. Data Mining Concepts and Techniques, Han, Kamber, 3rd Edition. 2. Student’s Handbook for Associate Analytics – II, III. 3. Data Science & Big Data Analytics: ICT ACADEMY by DELL EMC. 4. Mining of Massive Datasets, Jure Leskovec Stanford Univ. Anand Rajaraman Milliway Labs Jeffrey D Ullman Stanford Univ. | |
| E-Resources | 1. <https://nptel.ac.in/courses> | |