|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **DATA SCIENCE AND APPLICATIONS**  **(Effective from the Academic Year 2023 - 2024)**  **VI SEMESTER** | | | | | | | | | | | | | | | | | | | | | | | | | |
| Course Code | | | | | | | | | | **21DS62** | | | | | CIA Marks | | | | | | | | 50 | | |
| Number of Contact Hours/Week (L: T: P: S) | | | | | | | | | | 3:0:2:0 | | | | | SEE Marks | | | | | | | | 50 | | |
| Total Hours of Pedagogy | | | | | | | | | | 40L + 20P | | | | | Exam Hours | | | | | | | | 03 | | |
| **CREDITS – 4** | | | | | | | | | | | | | | | | | | | | | | | | | |
| **COURSE PREREQUISITES:**   * Fundamental knowledge of mathematical concepts, analytical skills and programming. | | | | | | | | | | | | | | | | | | | | | | | | | |
| **COURSE OBJECTIVES:**   * Develop relevant programming abilities. * Demonstrate proficiency with statistical analysis of data. * Develop the ability to build and assess data-based models. * Learn to execute statistical analyses. | | | | | | | | | | | | | | | | | | | | | | | | | |
| **TEACHING - LEARNING STRATEGY:**  Following are some sample strategies that can be incorporate for the Course Delivery   * Chalk and Talk Method/Blended Mode Method * Power Point Presentation * Expert Talk/Webinar/Seminar * Video Streaming/Self-Study/Simulations * Peer-to-Peer Activities * Activity/Problem Based Learning * Case Studies * MOOC/NPTEL Courses * Any other innovative initiatives with respect to the Course contents | | | | | | | | | | | | | | | | | | | | | | | | | |
| **COURSE CONTENTS** | | | | | | | | | | | | | | | | | | | | | | | | | |
| **MODULE - I** | | | | | | | | | | | | | | | | | | | | | | | | | |
| **Introduction**: The Ascendance of Data, What Is Data Science? , Visualizing Data, matplotlib, Bar Charts, Line Charts, Scatterplots, Statistics, Describing a Single Set of Data, Correlation, Simpson’s Paradox, Some Other Correlational Caveats, Correlation and Causation, Dependence and Independence, Conditional Probability, Bayes’s Theorem, Random Variables, Continuous Distributions, The Normal Distribution, The Central Limit Theorem.  ***Text book 1: Chapter 1, Chapter 3, Chapter 5 and Chapter 6*** | | | | | | | | | | | | | | | | | | | | | | | | **8 Hours** | |
| **MODULE - II** | | | | | | | | | | | | | | | | | | | | | | | | | |
| **Working with data:** Getting Data, stdin and stdout, Reading Files, Scraping the Web, Using APIs, Example: Using the Twitter APIs, Exploring the Data, Using NamedTuples, Dataclasses, Cleaning and Munging, Manipulating Data, Rescaling, An Aside: tqdm, Dimensionality Reduction. The Curse of Dimensionality, Main Approaches for Dimensionality Reduction, Projection, Manifold Learning, PCA, Preserving the Variance, Principal Components, Projecting Down to d Dimensions.  ***Text book 1: Chapter 9 and Chapter 10***  ***Text book 2: Chapter 8*** | | | | | | | | | | | | | | | | | | | | | | | | **8 Hours** | |
| **MODULE - III** | | | | | | | | | | | | | | | | | | | | | | | | | |
| **Training:** Models, Linear Regression, The Normal Equation, Computational Complexity, Gradient Descent, Batch Gradient Descent, Stochastic Gradient Descent, Mini-batch Gradient Descent, Polynomial Regression, Learning Curves, Regularized Linear Models , Ridge Regression, Lasso Regression, ElasticNet Early Stopping, Logistic Regression, Estimating Probabilities, Training and Cost Function Decision Boundaries, Softmax Regression  ***Text book 2: Chapter 4*** | | | | | | | | | | | | | | | | | | | | | | | | **8 Hours** | |
| **MODULE - IV** | | | | | | | | | | | | | | | | | | | | | | | | | |
| **Machine Learning and Deep Learning:** Machine Learning, Modeling, What Is Machine Learning, Overfitting and Underfitting, Correctness, The Bias-Variance Tradeoff, Feature Extraction and Selection, Neural Networks, Perceptrons, Feed-Forward Neural Networks, Backpropagation, Example: Fizz Buzz, Deep Learning, The Tensor, The Layer Abstraction, The Linear Layer, Neural Networks as a Sequence of Layers, Loss and Optimization, Other Activation Functions, Softmaxes and Cross-Entropy, Dropout, Saving and Loading Models, Clustering, The Idea, The Model, Choosing k, Bottom-Up Hierarchical Clustering.  ***Text book 1: Chapter 11, Chapter 18, and Chapter 19*** | | | | | | | | | | | | | | | | | | | | | | | | **8 Hours** | |
| **MODULE – V** | | | | | | | | | | | | | | | | | | | | | | | | | |
| **Network Analysis, Recommender System and MapReduce:** Betweenness Centrality, Eigenvector Centrality, Directed Graphs and PageRank, Recommender Systems, Manual Curation, Recommending What’s Popular, User-Based Collaborative Filtering, Item-Based Collaborative Filtering, Matrix Factorization, Create Table and Insert, Update, Delete, Select, Group By, Order By, Join, Subqueries, Indexes, Query Optimization, N0Sql, MapReduce, Example: Word count, Why MapReduce, MapReduce more generally, Example: Analyzing status updates, Example: Matrix multiplication, An Aside: Combiners.  ***Text book 1: Chapter 22, Chapter 23, Chapter 24 and Chapter 25*** | | | | | | | | | | | | | | | | | | | | | | | | **8 Hours** | |
| **COURSE OUTCOMES** | | | | | | | | | | | | | | | | | | | | | | | | | |
| Upon completion of this course, the students will be able to: | | | | | | | | | | | | | | | | | | | | | | | | | |
| **CO**  **No.** | **Course Outcome Description** | | | | | | | | | | | | | | | | | | | | | **Bloom’s Taxonomy Level** | | | |
| CO1 | Identify and demonstrate data using visualization tools and apply statistical analysis methods, to analyze and interpret data effectively in various real-world contexts. | | | | | | | | | | | | | | | | | | | | | CL3 | | | |
| CO2 | Apply techniques for data acquisition, exploring and preparing data, dimensionality reduction , to reduce dataset complexity while preserving key information. | | | | | | | | | | | | | | | | | | | | | CL3 | | | |
| CO3 | Apply gradient descent optimization techniques, to train linear regression models effectively. | | | | | | | | | | | | | | | | | | | | | CL3 | | | |
| CO4 | Demonstrate the use of machine learning and deep learning models to implement efficient data-driven solutions for real-world problems. | | | | | | | | | | | | | | | | | | | | | CL3 | | | |
| CO5 | Demonstrate knowledge about the recommender system, MapReduce and understand the importance of data ethics. | | | | | | | | | | | | | | | | | | | | | CL3 | | | |
| **LABORATORY COMPONENTS** | | | | | | | | | | | | | | | | | | | | | | | | | |
| **Exp. No.** | **Experiment Description** | | | | | | | | | | | | | | | | | | | **CO No.** | | **Bloom’s Taxonomy Level** | | | |
| 1 | Demonstrate all the basic plots using Matplotlib package and python programming. | | | | | | | | | | | | | | | | | | | CO1 | | CL3 | | | |
| 2 | Implement a python program to perform File Operations on Excel Dataset. | | | | | | | | | | | | | | | | | | | CO2 | | CL3 | | | |
| 3 | Write a python program to perform Array operations using the Numpy package. | | | | | | | | | | | | | | | | | | | CO2 | | CL3 | | | |
| 4 | Build an Artificial Neural Network by implementing the Backpropagation algorithm and test the same using appropriate data sets. | | | | | | | | | | | | | | | | | | | CO3 | | CL3 | | | |
| 5 | Demonstrate Linear Regression operation using python programming. | | | | | | | | | | | | | | | | | | | CO4 | | CL3 | | | |
| 6 | Train a regularized logistic regression classifier on the in-build iris dataset using scikit-learn. Train the model and report the best classification accuracy. | | | | | | | | | | | | | | | | | | | CO4 | | CL3 | | | |
| 7 | Write a python program to perform Data Manipulation operations using Pandas package. | | | | | | | | | | | | | | | | | | | CO4 | | CL3 | | | |
| 8 | Develop a MapReduce program to find the grades of students in python. | | | | | | | | | | | | | | | | | | | CO5 | | CL3 | | | |
| **CO-PO-PSO MAPPING** | | | | | | | | | | | | | | | | | | | | | | | | | |
| **CO**  **No.** | **Programme Outcomes (PO)** | | | | | | | | | | | | | | | | | | | | **Programme Specific Outcome (PSO)** | | | | |
| **1** | **2** | | **3** | **4** | | **5** | **6** | | | **7** | | **8** | **9** | | **10** | | **11** | **12** | | **1** | | | | **2** |
| **CO1** | 3 | 3 | | 3 | 2 | | 3 |  | | |  | |  |  | |  | | 2 | 2 | | 3 | | | | 2 |
| **CO2** | 3 | 3 | | 3 | 2 | | 3 |  | | |  | |  |  | |  | | 2 | 2 | | 3 | | | | 2 |
| **CO3** | 3 | 3 | | 3 | 2 | | 3 |  | | |  | |  |  | |  | | 2 | 2 | | 3 | | | | 2 |
| **CO4** | 3 | 3 | | 3 | 2 | | 3 |  | | |  | |  |  | |  | | 2 | 2 | | 3 | | | | 2 |
| **CO5** | 3 | 3 | | 3 | 2 | | 3 |  | | |  | |  |  | |  | | 2 | 2 | | 3 | | | | 2 |
| **3: Substantial (High)** | | | | | | | **2: Moderate (Medium)** | | | | | | | | | **1: Poor (Low)** | | | | | | | | | |
| **ASSESSMENT STRATEGY**  Assessment will be both CIA and SEE. Students learning will be assessed using Direct and Indirect methods: | | | | | | | | | | | | | | | | | | | | | | | | | |
| **Sl. No.** | **Assessment Description** | | | | | | | | | | | **Weightage (%)** | | | | | **Max. Marks** | | | | | | | | |
| **1** | **Continuous Internal Assessment (CIA)** | | | | | | | | | | | **100 %** | | | | | **50** | | | | | | | | |
|  | Continuous Internal Evaluation (CIE) | | | | | | | | | | | 60 % | | | | | 30 | | | | | | | | |
| Practical Session (Laboratory Component) | | | | | | | | | | | 40 % | | | | | 20 | | | | | | | | |
| **2** | **Semester End Examination (SEE)** | | | | | | | | | | | **100 %** | | | | | **50** | | | | | | | | |
| **ASSESSMENT DETAILS** | | | | | | | | | | | | | | | | | | | | | | | | | |
| **Continuous Internal Assessment (CIA) (50%)** | | | | | | | | | | | | | | | | | **Semester End Exam (SEE) (50%)** | | | | | | | | |
| **Continuous Internal Evaluation (CIE) (60%)** | | | | | | | | | **Practical Sessions (40%)** | | | | | | | |
| **I** | | | **II** | | | **III** | | |
| **Syllabus Coverage** | | | | | | | | | **Syllabus Coverage** | | | | | | | | **Syllabus Coverage** | | | | | | | | |
| **40%** | | | **30%** | | | **30%** | | | **100%** | | | | | | | | **100%** | | | | | | | | |
| MI | | |  | | |  | | | MI | | | | | | | | MI | | | | | | | | |
| MII | | | MII | | |  | | | MII | | | | | | | | MII | | | | | | | | |
|  | | | MIII | | |  | | | MIII | | | | | | | | MIII | | | | | | | | |
|  | | |  | | | MIV | | | MIV | | | | | | | | MIV | | | | | | | | |
|  | | |  | | | MV | | | MV | | | | | | | | MV | | | | | | | | |
| **NOTE:**   * Assessment will be both CIA and SEE. * The practical sessions of the IPCC shall be for CIE only. * The Theory component of the IPCC shall be for both CIA and SEE respectively. * The questions from the practical sessions shall be included in Theory SEE. | | | | | | | | | | | | | | | | | | | | | | | | | |
| ***Note: For Examinations (both CIE and SEE), the question papers shall contain the questions mapped to the appropriate Bloom’s Level. Any COs mapped with higher cognitive Bloom’s Level may also be assessed through the assignments.*** | | | | | | | | | | | | | | | | | | | | | | | | | |
| **SEE QUESTION PAPER PATTERN:**   1. The question paper will have **TEN** full questions from **FIVE** Modules 2. There will be 2 full questions from each module. Every question will carry a maximum of 20 marks. 3. Each full question may have a maximum of four sub-questions covering all the topics under a module. 4. The students will have to answer FIVE full questions, selecting one full question from each module. | | | | | | | | | | | | | | | | | | | | | | | | | |
| **TEXT BOOKS:**   1. Joel Grus, Data Science from Scratch-First Principles with Python, O’ Reilly Publications, 2nd Edition, 2019, ISBN: 978-9352138326. 2. Aurelien Geron, Hands-on Machine Learning with Scikit-Learn & TensorFlow, O’Reilly Media Publications, 3rd Edition, 2022, ISBN: 978-93-5542-198-2. | | | | | | | | | | | | | | | | | | | | | | | | | |
| **REFERENCE BOOKS:**   1. Emily Robinson, Jacqueline Nolis, Build a Career in Data Science, Manning Publications, 1st Edition, 2020, ISBN: 9781638350156. 2. Aurelien Geron, Hands-on Machine Learning with Scikit-Learn & TensorFlow, O’Reilly Media Publications, 3rd Edition, 2022, ISBN: 978-93-5542-198-2. | | | | | | | | | | | | | | | | | | | | | | | | | |
| **REFERENCE WEB LINKS AND VIDEO LECTURES (E - RESOURCES):**   1. https://onlinecourses.nptel.ac.in/noc21\_cs69 2. https://onlinecourses.nptel.ac.in/noc22\_cs32 | | | | | | | | | | | | | | | | | | | | | | | | | |