INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR

1. **Name of the Academic Unit:** Centre of Excellence in Artificial Intelligence

2. **Subject Name:** Introduction to Data Science
   L-T-P: 3-1-0
   Credits: 4

3. **Pre-requisites:** None

4. **Syllabus and reference books:**

Syllabus:

The aim of this subject is to give the students their first experience of handling real-world data, and equip them with the necessary mathematical and statistical tools for the same purpose. By the end of the semester, the students should have learnt the nature and characteristics of real-world data from different domains, and tools and techniques to explore and visualize them.

Introduction to data types, data representation, exploratory data analysis, hypothesis testing, modes of data, basic predictive models, case studies

Reference Books:

1) **Essential Math for Data Science** by Thomas Nield, O'Reilly Media, Inc. (May 2022)

2) **Practical Statistics for Data Scientists** by Peter Bruce, Andrew Bruce, O'Reilly Media, Inc (May 2017)

3) **The Art of Data Science** by Roger D. Peng and Elizabeth Matsui, lulu.com (2016)

5. Lecture-wise break-up:

SL No	Module	Topic	No. of lecture-hours (48)
1	**Introduction to Data Types**	Collection of data: manual, instrument-based Types of data: textual, numeric, audio/visual	**6 hours**
2	**Data Representation**	Data visualization using charts, graphs and histograms, Vectors, Matrices and Random Variables, with associated concepts of probability and statistics	**8 hours**
3	**Exploratory Data Analysis**	univariate, bivariate analysis, correlation analysis, rank correlation Pre-processing: i) imputing missing values, ii) identify anomalies Time-series data: extraction of trends, periodicity etc Decomposition techniques: PCA, EMD, t-SNE, Multidimensional scaling	**10 hours**
4	**Hypothesis Testing**	Statistical tests, 2-sample and permutation tests, chi-square tests and ANOVA	**6 hours**
5	**Modes of Data**	Image data analysis: basic denoising, feature-based image representation, entity-based image representation Text data analysis: tokenization, lemmatization, stemming, stop-word removal, TF-IDF Speech data analysis:	**10 hours**
6	**Basic Predictive Models**	Linear Regression, Nearest-Neighbor, ARIMA type models, DB-Scan	**4 hours**
7	**Miscellaneous**	Use-cases and short projects	**4 hours**