Professional Electives: Computational & Data Science Stream

DATA SCIENCE

[As per Choice Based Credit System (CBCS) scheme]

SEMESTER – III

Course Code	:	23CSE5107	**Credits**	:	04
Hours / Week	:	03 Hours	**Total Hours**	:	39 (T) +26(P) Hrs
L–T–P–J	:	3–0–2–0			

Course Learning Objectives:

This Course will enable students to:

1. **Apply** processes suitable to data preprocessing and transformation to be able to prepare data to extract insights.
2. **Visualize** data by computing and display graphs and plots to identify relationships and patterns and by modelling exploratory data analytics.
3. **Utilize** mathematical and statistical techniques to test hypothesis.
4. **Employ** central limit theorem and confidence interval enabling them to model real-world phenomena and make accurate predictions**.**
5. **Use** open-source tools to engage in practical application of the data to formulate problem statements and work to identify solutions.

Teaching-Learning Process (General Instructions)

These are sample new pedagogical methods, where teacher can use to accelerate the attainment of the various course outcomes.

1. ***Lecture method*** means it includes not only traditional lecture method, but different *type of teaching methods* may be adopted to develop the course outcomes.
2. ***Interactive Teaching:*** *Adopt the* ***Active learning*** that includes brainstorming, discussing, group work, focused listening, formulating questions, notetaking, annotating, and roleplaying.
3. Show ***Video/animation*** films to explain functioning of various concepts.
4. Encourage ***Collaborative*** (Group Learning) Learning in the class.
5. To make ***Critical thinking***, ask at least three Higher order Thinking questions in the class.
6. Adopt ***Problem Based Learning***, which fosters students' Analytical skills, develop thinking skills such as the ability to evaluate, generalize, and analyse information rather than simply recall it.
7. Show the ***different ways to solve*** the same problem and encourage the students to come up with their own creative ways to solve them.
8. Discuss how every ***concept can be applied to the real world*** - and when that's possible, it helps improve the students' understanding.

UNIT – I: About Data	**09 Hours**
Introduction, Causality and Experiments - Data Pre-processing: Knowing data, Data cleaning, Data reduction, Data transformation, Data discretization.	
UNIT – II: Data Visualization	**09 Hours**
Visualization and Graphing: Visualizing Categorical Distributions, Visualizing Numerical Distributions, Overlaid Graphs, plots, and summary statistics of Exploratory Data Analysis (EDA). Exploring Univariate Data - Histograms -Stem-and-Leaf Quantile Based Plots - Continuous Distributions -Quantile Plots- QQ Plot- Box Plots.	
UNIT – III: Statistics	**06 Hours**
Introduction to Statistics- Sampling, Sample Means and Sample variance sample moments, covariance, correlation, Sampling Distributions - Parameter Estimation Bias -Mean Squared Error -Relative Efficiency – Standard Error - Maximum Likelihood Estimation. Empirical Distributions- Sampling from a Population- Empirical Distribution of a Statistic -Testing Hypotheses Error probabilities- Assessing Models-Multiple Categories -Decisions and Uncertainty- Comparing Two Samples -A/B Testing - ANOVA.	
UNIT – IV: Sampling theory	**07 Hours**
Estimation- Percentiles- The Bootstrap - Confidence Intervals- Using Confidence Intervals - The SD and the Normal Curve - The Central Limit Theorem - point and interval estimation, Prediction- Correlation -The Regression Line -The Method of Least Squares - Least Squares	

UNIT – V: Case studies on using computational tools for data analytics	08 Hours
Case studies on Visualization with the help of Tools like Altair Tableau, Rapid miner, and MATLAB. [Access to open-source tools will be granted for practical application to work on cases studies]	

Course Outcome	Description	Bloom's Taxonomy Level
At the end of the course the student will be able to:		
1	**Apply** their knowledge of data preprocessing and transformation to be able to prepare data to extract insights.	L3
2	**Visualize** data by creating graphs and plots to identify relationships and patterns and by modelling exploratory data analytics.	L3
3	**Utilize** mathematical and statistical techniques to test hypothesis and to identify covariance with A/B testing and Analysis of Variance.	L3
4	**Employ** central limit theorem and confidence interval enabling them to model real-world phenomena and make accurate predictions**.**	L2 & L3
5	**Use** open-source tools to engage in practical application of the data to formulate problem statements and work to identify solutions and to build models.	L4

Table: Mapping Levels of COs to POs / PSOs

COs	Program Outcomes (POs)												PSOs	
	1	2	3	4	5	6	7	8	9	10	11	12	1	2
CO1	3	2	2	1					1					
CO2	3	2	2						1					
CO3	3	2	2	2					1					
CO4	3	2	2	2					1					
CO5	3	3	2	1	2				1				2	

3: Substantial (High) **2: Moderate (Medium)** **1: Poor (Low)**

TEXT BOOKS:

1. Adi Adhikari and John De Nero, "Computational and Inferential Thinking: The Foundations of Data Science", 1st edition, 2019.
2. Jiawei Han, Micheline Kamber, Jian Pei, "Data Mining Concepts and Techniques", 4th Edition, Elsevier, 2006.
3. Douglas C. Montgomery, George C. Runger, "Applied Statistics and Probability for Engineers", 6th Edition, Wiley,2013.

REFERENCE BOOKS:

1. Wendy L. Martinez, Angel R. Martinez, "Computational Statistics Handbook with MATLAB", 2nd Edition, Chapman Hall/CRC, 2008.

E-Resources:

1. Data Science for Engineers, IIT Madras- https://nptel.ac.in/courses/106106179
2. https://ifacet.iitk.ac.in/professional-certificate-course-in-data-science/
3. https://online.stat.psu.edu/stat506/lesson/1/1.4
4. https://onlinestatbook.com/2/advanced_graphs/q-q_plots.html

****