STATISTICAL FOUNDATIONS OF DATA SCIENCE

V Semester: **CSE (AI & ML)**

Course Code	Category	Hours / Week			Credits	Maximum Marks		
		L	T	P	C	CIA	SEE	Total
ACAC07	Elective	3	0	0	3	30	70	100
Contact Classes: 45	Tutorial Classes: Nil	Practical Classes: Nil				Total Classes: 45		

Prerequisites: **Linear Algebra and Calculus, Probability and Statistics**

I. COURSE OVERVIEW:

The course is designed to introduce to the basics of data science, graphics, and modeling. Topics covered include flavors of data, basic mathematics, probability and statistics and data visualization. The main objective of the course is to teach a range of topics and concepts related to the data science process. This course reaches to student by power point presentations, lecture notes, and lab which will give you the chance to apply knowledge of data science process. This course is very helpful forthe artificial intelligence techniques.

II. COURSE OBJECTIVES:

The students will try to learn:

I. The fundamental knowledge on basics of data science.
II. The basic principles of data acquisition, exploring and modeling data efficiently.
III. The foundations of probability and statistics for data science.
IV. The current scope, potential applications of data science.

III. COURSE OUTCOMES:

After successful completion of the course, students should be able to:

CO 1	**Recall** the categories and levels of data using steps involved indata science.	Remember
CO 2	**Demonstrate** the data pre-processing terms for improving the quality of dataset using processes such as feature generation andfeature selection	Understand
CO 3	**Solve** mathematical problems using various arithmetic and more challenging forms of math.	Apply
CO 4	**Apply** probability theorems and approaches for calculating the number of outcomes of the events.	Apply
CO 5	**Illustrate** the obtaining and sampling data in statistics to quantify and visualize our data.	Understand
CO 6	**Summarize** the concepts of communication by using the visualization and presenting strategies.	Understand

IV. SYLLABUS:

MODULE – I: FLAVORS OF DATA (09)

Flavors of Data: Structured versus unstructured data, Quantitative and qualitative data, The four levels of data: Nominal level, Ordinal level, Interval level, and Ratio level, The five steps of Data Science: Ask an interesting question, obtain the data, explore the data, model the data, communicate and visualize the results, Explore the data.

MODULE – II: DATA PRE-PROCESSING AND FEATURE SELECTION (09)

Data Pre-processing: Data cleaning, Data integration, Data Reduction, Data Transformation and Data Discretization, Feature Generation and Feature Selection, Feature Selection algorithms: Filters, Wrappers, Decision Trees, Random Forests.

MODULE – III: BASIC MATHEMATICS AND PROBABILITY FOR DATA SCIENCE (09)

Mathematics: Vectors and matrices, Arithmetic symbols, Graphs, Logarithms/exponents, Set theory, Linear algebra.
Probability: Basic definitions, Probability, Bayesian versus Frequentist, Compound events, Conditional Probability, The rules of probability, Collectively exhaustive events, Bayes theorem, Random variables.

MODULE – IV: STATISTICS FOR DATA SCIENCE (09)

Statistics: Obtaining data, Sampling data, Measuring Statistics, The Empirical rule, Point estimates, Sampling distributions, Confidence intervals, Hypothesis tests.

MODULE – V: COMMUNICATING DATA (09)

Data Visualization: Identifying effective and ineffective visualizations: Scatter plots, Line graphs, Bar charts, Histograms, Box plots. Graphs and Statistics lie: Correlation versus causation, Simpson's paradox, Verbal Communication, The why/how/what strategy of presenting.

V. TEXT BOOKS:

1. Sinan Ozdemir, "Principles of Data Science: Learn the techniques and math you need to start making sense of your data", 1st edition, Packt publishing, 2016.
2. Jianqing Fan, Runze Li, Cun-Hui Zhang, Hui Zou, "Statistical Foundations of Data Science", Chapman and Hall / CRC Press, 2020.

VI. REFERENCE BOOKS:

1. Cathy O'Neil, Rachel Schutt, "Doing Data Science: Straight talk from the frontline", 1st edition, O'Reilly 2014.
2. James G, Witten D, Hastie T, Tibshirani R, "An Introduction to Statistical Learning with applications in R", Springer, 2013.
3. Hastie Trevor, Tibshirani Robert, Friedman Jerome, "The Elements of Statistical Learning Data Mining, Inference and Prediction", 2nd edition, 2009.

VII. WEB REFERENCES:

1. https://www.analyticsvidhya.com/blog/tag/statistics-for-data-science/
2. https://towardsdatascience.com/fundamentals-of-statistics-for-data-scientists-and-data-analysts-69d93a05aae7
3. https://fan.princeton.edu/fan/classes/525/TableOfContent.pdf
4. https://www.stat.berkeley.edu/~mmahoney/talks/foundations_apr16.pdf
5. https://nptel.ac.in/courses/106/106/106106179/