INSTITUTE OF AERONAUTICAL ENGINEERING
(Autonomous)
Dundigal - 500 043, Hyderabad, Telangana

COURSE CONTENT

STATISTICAL FOUNDATIONS OF DATA SCIENCE								
V Semester: CSE (AI & ML)								
Course Code	**Category**	**Hours / Week**			**Credits**	**Maximum Marks**		
ACAD07	**Elective**	**L**	**T**	**P**	**C**	**CIA**	**SEE**	**Total**
		3	0	0	3	40	60	100
Contact Classes: 48	**Tutorial Classes: NIL**	**Practical Classes: Nil**				**Total Classes: 48**		
Prerequisite: 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 for the artificial intelligence techniques.

II.COURSES OBJECTIVES:

The students will try to learn:

I. The fundamental types and levels of data, and how to formulate meaningful questions to guide data analysis through the key stages of the data science lifecycle.
II. Essential data preprocessing, feature selection, and mathematical foundations necessary for handling, preparing, and modeling real-world datasets.
III. Core statistical reasoning and effective data communication skills to analyze, visualize, and clearly present data-driven insights.

III. COURSE OUTCOMES:

At the end of the course, students should be able to:

CO1	Identify and differentiate between various types and levels of data used in data science.
CO2	Apply data preprocessing techniques such as cleaning, transformation, and feature selection to prepare datasets.
CO3	Use mathematical and probabilistic concepts to model data and support analytical tasks.
CO4	Perform statistical analysis including hypothesis testing, estimation, and interpretation of sampling distributions.
CO5	Create effective visualizations using plots and charts to interpret and present data insights.
CO6	Communicate data-driven findings using structured verbal and visual storytelling approaches.

IV.COURSE CONTENT:

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", 1 st 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", 1 st 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/

VIII. MATERIAL ONLINE:

1. Course template
2. Tutorial question bank
3. Tech talk topics
4. Open-ended experiments
5. Definitions and terminology
6. Assignments
7. Model question paper – I
8. Model question paper – II
9. Lecture notes
10. PowerPoint presentation
11. E-Learning Readiness Videos (ELRV)