DSPC3001 DATA SCIENCE (3-0-0)

Course Objectives:

- To understand and apply basic statistical and data analysis techniques.
- To preprocess and visualize real-world data.
- To equip students with practical skills to prepare data for analysis using real-world datasets.
- To foster proficiency in data science tools and programming.
- To enable students to perform exploratory data analysis (EDA).
- To communicate results effectively with visualizations.

Module-I: (08 Hrs)

Introduction to data science:

Definition and scope of data science, Data science process, key components, roles; Data Science Project Life Cycle: OSEMN Framework; Tools and technologies such as Python, R, Git etc.

Data Collection and preprocessing:

Types of data: structured, semi-structured, unstructured; Sources of data: web scraping, APIs, databases; Data cleaning: handling missing values, outliers, noise; Data transformation: normalization, encoding categorical data.

Module-II: (08 Hrs)

Statistical Foundations for Data Science:

Types of data: nominal, ordinal, discrete, continuous; Descriptive statistics: mean, median, mode, variance, standard deviation; Probability theory: basic rules, conditional probability, Bayes' theorem; Probability distributions: Normal, Binomial, Poisson, Exponential; Inferential statistics: sampling, hypothesis testing, p-values, confidence intervals; Chi-square test, t-test, ANOVA.

Module-III: (08 Hrs)

Data Collection, Preparation, and Cleaning:

Data sources: databases, APIs, web scraping; Data formats: CSV, JSON, XML, Excel, SQL; Data wrangling: missing value treatment, outlier detection; Data transformation: scaling, encoding categorical variables, feature extraction; Data integration and schema mapping; Tools: Pandas (Python), dplyr (R), NumPy, Scikit-learn, Matplotlib, Seaborn.

Module-IV: (08 Hrs)

Exploratory Data Analysis (EDA):

Definition, importance, Types-Univariate, bivariate and multivariate analysis; steps for performing EDA; Correlation and covariance; Data visualization techniques: Histograms, bar charts, box plots, scatter plots, heatmaps; Pair plots and joint plots; Dimensionality reduction: PCA (Principal Component Analysis) – basics only; Libraries: Matplotlib, Seaborn, Plotly

Module-V: (08 Hrs)

Introduction to Big Data and Cloud in Data Science:

Overview of Big Data concepts: The four dimensions of Big Data: volume, velocity, variety, veracity; Distributed Hash-table, Key-Value Storage Model (Amazon's Dynamo), Document Storage Model (Facebook's Cassandra), Graph storage models; Hadoop Ecosystem: HDFS, Graph Representation in MapReduce: Graph Processing with Spark, Spark GraphX, GraphX features, Examples, Graph Algorithms-Shortest Path Algorithm; Apache Spark basics; Introduction to cloud platforms: AWS, Azure, Google Cloud for data science; Role of cloud in scalable data processing and deployment.

Course Outcomes:

- At the end of the course the student will be able to:
- Understand the fundamental concepts and scope of Data Science.
- Apply statistical and mathematical methods to analyze data.
- Collect, clean, and preprocess raw data for analysis
- Perform exploratory data analysis (EDA) and visualization
- Use programming tools and libraries for data analysis
- Understand the basics of big data processing and cloud technologies

Text Books:

1. Gareth James Daniela Witten Trevor Hastie, Robert Tibshirani, An Introduction to Statistical Learning with Applications in R.
2. Han, Kamber, and J Pei, Data Mining Concepts and Techniques.
3. Understanding Big Data Analytics for Enterprise Class Hadoop and Streaming Data by Dirk deRoos, Chris Eaton, George Lapis, Paul Zikopoulos, Tom Deutsch, 1st Edition, TMH,2012.
4. Learning Spark “Holden KarauA. Konwinskietc.,” O’Reilly Publications.

Reference Books:

1. C Bishop, "Pattern Recognition and Machine Learning", Springer, 2006.
2. Chun-houh Chen, Wolfgang Hardle, Antony Unwin, Handbook of Data Visualization, Springer, 2008.