Course code	Course Title	L	T	P	C
BCSE331L	Exploratory Data Analysis	2	0	0	2
Pre-requisite	NIL	Syllabus version			
		1.0			

Course Objectives

1. The course introduces the methods for data preparation and data understanding.
2. It covers essential exploratory techniques for understanding multivariate data by summarizing it through statistical and graphical methods.
3. Supports to summarize use of predictive analytics, data science and data visualization.

Course Outcomes

At the end of the course, the student will be able to

1. Handle missing data in the real world data sets by choosing appropriate methods.
2. Summarize the data using basic statistics. Visualize the data using basic graphs and plots.
3. Identify the outliers if any in the data set.
4. Choose appropriate feature selection and dimensionality reduction.
5. Apply Techniques for handling multi-dimensional data.

Module:1	Introduction to Exploratory Data Analysis	4 hours

Introduction to Exploratory Data Analysis (EDA) –Steps in EDA, Data Types: Numerical Data – Discrete data, continuous data – Categorical data – Measurement Scales: Nominal, Ordinal, Interval, Ratio – Comparing EDA with classical and Bayesian Analysis – Software tools for EDA.

Module:2	Data Transformation	4 hours

Transformation Techniques: Performing data deduplication - replacing values – Discretization and binning. Introduction to Missing data, handling missing data: Traditional methods - Maximum Likelihood Estimation.

Module:3	Correlation Analysis and Time Series Analysis	4 hours

Types of analysis: Univariate analysis - bivariate analysis - multivariate analysis. Time Series Analysis (TSA): Fundamentals of TSA - characteristics of TSA – Time based indexing - visualizing time series – grouping time series data - resampling time series data.

Module:4	Data Summarization and Visualization	4 hours

Statistical summary measures, data elaboration, 1-D Statistical data analysis, 2-D Statistical data Analysis, contingency tables, n-D Statistical data analysis. Visualization: Scatter plots – Dot charts - Bar plots.

Module:5	Clustering Algorithms	4 hours

Introduction to Spectral clustering – Document clustering – Minimum Spanning Tree clustering. Overview of Model-based clustering – Expectation-Maximization algorithm – Hierarchical Agglomerative model-based clustering. Outlier detection using Clustering.

Module:6	Dimensionality Reduction	4 hours

Linear Methods: Principal Component Analysis (PCA) – Singular Value Decomposition – Factor Analysis -Intrinsic Dimensionality. Non Linear methods: Multidimensional Scaling – Manifold Learning – Self-Organizing Maps.

Module:7	Model Development and Evaluation	4 hours

Constructing linear regression model – evaluation – computing accuracy – understanding accuracy. Understanding reinforcement learning: Difference between supervised and reinforcement learning – Applications of reinforcement learning.

Module:8	Contemporary Issues	2 hours

	Total Lecture hours:	**30hours**

Text Book(s)	
1.	Suresh Kumar Mukhiya, Usman Ahmed, "Hands-On Exploratory Data Analysis with Python" 1st Edition, 2020, Packt Publishing.
2.	Martinez, W , Martinez A & J.L. Solka : Exploratory Data Analysis with MATLAB, CRC Press, A Chapman & Hall Book, 3rd Edition, 2017
Reference Books	
1.	Michael Jambu, "Exploratory and multivariate data analysis", 1991, 1st Edition, Academic Press Inc.
2.	Charu C. Aggarwal, "Data Mining The Text book", 2015, Springer.
3.	Craig K. Enders, "Applied Missing Data Analysis", 2010, 1st Edition, The Guilford Press.
Mode of Evaluation: CAT / written assignment / Quiz / FAT / Project	

Recommended by Board of Studies	12-05-2022		
Approved by Academic Council	No. 66	Date	16-06-2022