

Data Engineering and Analytics	L	T	P	Probability, Statistics, Linear Algebra
	3	0	2	

Course Objective: 1. To familiarize with the concepts of data mining, data storage, data pipeline and data integration.
2.To analyze data using classification, clustering, feature selection, association rule mining and itemset mining.
3. To understand and explore data warehousing architectures and data lakes.

S. NO.	Course Outcomes (CO)
CO1	Ability to understand and describe data mining, data storage, data pipeline and data integration concepts.
CO2	Ability to understand and implement classification and clustering algorithms for analyzing linearly and non-linearly separable data.
CO3	Ability to understand and implement feature selection algorithms for data reduction.
CO4	Ability to understand and explore patterns and dependencies in data using association-rule mining and itemset-mining.
CO5	Ability to understand, explore and describe data warehousing architectures and data lakes.

S. NO.	Contents	Contact Hours
UNIT 1	Introduction to data engineering: Storing data, Data loading, Data transformation, Data structures, SQL and NoSQL databases, Database normalization, Data cubes, Snowflake scheme, Data warehouses, Data lakes, Data Marts, Metadata in Data Warehouse, Data pipeline, Data integration- Extract, Transform, and Load (ETL) processes for integrating data from multiple sources.	10
UNIT 2	Advanced data analytics and machine learning: Cluster Analysis, Types of Data in Cluster Analysis, Partitioning methods, Hierarchical Methods. Classification techniques for linearly separable and non-linearly separable data- Linear Discriminant Analysis and Support Vector Machine. Feature Ranking and Feature Selection Algorithms. Associations and correlations- basic concepts, efficient and scalable frequent item sets mining methods, mining various kinds of association rules, constraint-based association mining.	12
UNIT 3	Data warehousing: Planning Your Data Warehouse, The Data Warehouse Project, Architectural Components: Understanding Data Warehouse Architecture, Infrastructure Supporting Architecture, Collection of Tools. Indexing the data warehouse, performance enhancement techniques.	10

UNIT 4	Data design and data preparation: From Requirements to Data Design, The STAR Schema, STAR Schema Keys, Advantages of the STAR Schema. Data modeling for data warehouses. Scalable data engineering solutions - handling massive datasets. Online Analytical Processing models, Online Transaction Processing models. .	10
	TOTAL	42