Gyanmanjari
Innovative University

Course Syllabus
Gyanmanjari Institute of Technology
Semester-6 (B. Tech.)

Subject: Data Warehousing and Mining-BETIT16325

Type of course: Professional Core

Prerequisite: Students should have a basic understanding of database management systems (DBMS), including concepts of relational databases, SQL queries, and normalization. Familiarity with data structures, algorithms, and fundamental concepts of statistics and probability is also expected. Basic knowledge of programming (preferably in Python or Java) will help in implementing data mining algorithms and working with tools for data analysis.

Rationale:

In the era of information explosion, organizations require efficient methods to store, manage, and extract meaningful knowledge from vast amounts of data. Data Warehousing provides the foundation for integrating heterogeneous data into a unified repository for analysis, while Data Mining offers techniques to discover hidden patterns, correlations, and trends that support decision-making. This course equips students with theoretical understanding and practical skills in OLAP, data preprocessing, classification, clustering, and association analysis. It bridges database concepts with modern analytical techniques, enabling students to apply advanced tools in real-world domains such as business intelligence, healthcare, finance, and e-commerce.

Teaching and Examination Scheme:

Teaching Scheme			Credits	Examination Marks					Total Marks
CI	T	P	C	Theory Marks		Practical Marks		CA	
				ESE	MSE	V	P	ALA	
4	0	2	5	60	30	10	20	30	150

Legends: CI-ClassRoom Instructions; T – Tutorial; P - Practical; C – Credit; ESE - End Semester Examination; MSE- Mid Semester Examination; V – Viva; CA - Continuous Assessment; ALA- Active Learning Activities.

GYANMANJARI INNOVATIVE UNIVERSITY GMIU BHAVNAGAR

Course Content:

Sr. No	Course Content	Hrs.	% Weightage
1	**Fundamentals of Data Warehousing:** Evolution of decision support systems, characteristics of a data warehouse, OLTP vs OLAP systems. Data warehouse architecture: top-down, bottom-up, hybrid approaches. Multidimensional data model: fact tables, dimension tables, schemas (Star, Snowflake, Fact Constellation). ETL process: extraction, cleaning, transformation, loading. Data preprocessing: data integration, data reduction, data transformation, discretization.	12	20%
2	**Data Warehouse Implementation and OLAP:** OLAP models: MOLAP, ROLAP, HOLAP. OLAP operations (roll-up, drill-down, slice, dice, pivot). Data cube computation, indexing and materialized view selection. Query optimization techniques in data warehouses. Tools and platforms for OLAP and business intelligence (case studies).	12	20%
3	**Introduction to Data Mining:** Knowledge discovery in databases (KDD) process. Types of data mining tasks: descriptive vs predictive. Data mining functionalities: concept description, classification, clustering, association, outlier analysis. Challenges in data mining: scalability, high dimensionality, noisy/incomplete data, distributed and heterogeneous databases. Applications in business, scientific, and social domains.	12	20%
4	**Classification and Prediction Techniques:** Decision tree algorithms: ID3, C4.5, CART. Bayesian classifiers: Naïve Bayes, Bayesian Belief Networks. Instance-based learning: K-Nearest Neighbor (KNN). Regression: linear regression, logistic regression. Model evaluation and validation: confusion matrix, accuracy, precision, recall, F1-score, ROC curve, k-fold cross-validation, bootstrapping. Ensemble methods: Bagging, Boosting, Random Forest (overview).	12	20%
5	**Clustering, Association & Advanced Topics:** Clustering: Partitioning methods (K-Means, K-Medoids), Hierarchical clustering (agglomerative, divisive), Density-based methods (DBSCAN), Cluster evaluation and validity indices. Association rule mining: Apriori algorithm, FP-Growth, measures of association (support, confidence, lift, conviction). Advanced topics: Web mining (content, structure, usage mining), Text mining (text preprocessing, term frequency, sentiment analysis), Introduction to Big Data mining (Hadoop, Spark). Real-world case studies (retail, healthcare, finance).	12	20%

Continuous Assessment:

Sr. No	Active Learning Activities	Marks
1	**Enterprise Data Warehouse & OLAP Design Project:** Student will design an enterprise-level data warehouse for a chosen domain (e.g., banking fraud detection, e-commerce sales, or hospital analytics) and complete the task individually. The student must implement a multidimensional schema (Star/Snowflake/Fact Constellation), perform ETL on a moderately large dataset, and demonstrate OLAP queries (roll-up, drill-down, slice, dice) with proper result interpretation. Final submission includes the schema, ETL workflow, OLAP queries, and an insights report uploaded to the GMIU Portal.	10
2	**Advanced Association & Sequence Pattern Mining:** Student will implement association rule mining techniques (Apriori / FP-Growth) and extend the work to sequence pattern mining, using a transactional dataset (e.g., clickstream data or retail purchase data), and complete the task individually. The student must compare the generated rules using support, confidence, lift, and conviction, and interpret their business value. Final submission includes the code, experimental results, comparative analysis, and business recommendations uploaded to the GMIU Portal.	10
3	**Integrated Mining Challenge – Clustering vs Classification:** Student will run both clustering techniques (K-Means, DBSCAN) and classification algorithms (Decision Tree, Naïve Bayes, or Random Forest) on a real-world dataset (healthcare, finance, or social media) and complete the task individually. The student must compare unsupervised and supervised learning outcomes, evaluate performance metrics (precision, recall, F1-score), and discuss which approach provides better decision support. Final submission includes the source code, visualization outputs, and an evaluation report uploaded to the GMIU Portal.	10
	Total	30

Suggested Specification table with Marks (Theory): 60

Distribution of Theory Marks **(Revised Bloom's Taxonomy)**						
Level	**Remembrance (R)**	**Understanding (U)**	**Application (A)**	**Analyze (N)**	**Evaluate (E)**	**Create (C)**
Weightage %	25%	25%	15%	15%	10%	10%

Course Outcome:

After learning the course, the students should be able to:	
CO1	Understand the architecture and concepts of data warehouses, schemas, and OLAP operations for decision support.
CO2	Apply ETL processes and data preprocessing techniques for preparing clean and integrated datasets.
CO3	Analyze datasets using classification and prediction models such as Decision Trees, Naïve Bayes, KNN, and regression.
CO4	Implement clustering and association rule mining techniques (K-Means, DBSCAN, Apriori, FP-Growth) and evaluate their effectiveness.
CO5	Evaluate and interpret the application of data warehousing and data mining techniques in real-world domains such as business intelligence, healthcare, finance, and e-commerce.

List of Practical

Sr. No	Description	Unit No	Hrs.
1	Design and implement Star and Snowflake schemas for a given domain dataset (e.g., retail sales, hospital management, or university records) to demonstrate fact and dimension tables.	01	02
2	Perform an ETL process (Extraction, Transformation, and Loading) on a sample dataset using SQL/ETL tools. Students will clean, transform, and load the data into a warehouse schema.	01	02
3	Implement and execute OLAP operations (roll-up, drill-down, slice, dice, and pivot) on a data warehouse. Students will run queries and analyze multidimensional data.	01	04
4	Apply data preprocessing techniques such as missing value handling, normalization, integration of multiple datasets, and data reduction for mining readiness.	02	02
5	Implement a Decision Tree classifier (ID3 or C4.5) on a dataset and visualize the generated rules. Interpret the decision paths for classification tasks.	02	02
6	Apply Naïve Bayes classification on a dataset (e.g., text classification, student performance) and evaluate model accuracy using confusion matrix, precision, recall, and F1-score.	03	02
7	Implement K-Nearest Neighbor (KNN) classification and compare its performance with Decision Tree and Naïve Bayes using accuracy and error metrics.	03	02

GYANMANJARI INNOVATIVE UNIVERSITY BHAVNAGAR

8	Perform Linear Regression and Logistic Regression for prediction tasks such as sales forecasting or medical diagnosis, and evaluate performance using error measures.	03	02
9	Apply K-Means clustering on a dataset (e.g., customer segmentation, student performance) and visualize the resulting clusters with interpretation of result.	04	02
10	Implement DBSCAN or Hierarchical Clustering and compare the quality of clusters with K-Means using cluster evaluation indices.	04	02
11	Perform Association Rule Mining using Apriori algorithm on a transactional dataset (e.g., market basket analysis). Generate rules and calculate support, confidence, and lift.	05	04
12	Implement FP-Growth algorithm for frequent itemset mining and compare its performance and scalability with Apriori.	05	04
		Total	30

Instructional Method:

The course delivery method will depend upon the requirement of content and needs of students. The teacher in addition to conventional teaching method by black board, may also use any of tools such as demonstration, role play, Quiz, brainstorming, MOOCs etc.

From the content 10% topics are suggested for flipped mode instruction.

Students will use supplementary resources such as online videos, NPTEL/SWAYAM videos, e-courses, Virtual Laboratory.

The internal evaluation will be done based on the Active Learning Assignment.

Practical/Viva examination will be conducted at the end of semester for evaluation of performance of students in laboratory.

Reference Books:

[1] Jiawei Han, Micheline Kamber, Jian Pei – Data Mining: Concepts and Techniques – Morgan Kaufmann, 4th Edition, 2022.
[2] Alex Berson, Stephen J. Smith – Data Warehousing, Data Mining, and OLAP – McGraw-Hill, 2008.
[3] Margaret H. Dunham, S. Sridhar – Data Mining: Introductory and Advanced Topics – Pearson, 2nd Edition, 2015.
[4] Ralph Kimball, Margy Ross – The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling – Wiley, 3rd Edition, 2013.
[5] Arun K. Pujari – Data Mining Techniques – Universities Press, 2nd Edition, 2013.