CSI3010	DATA WAREHOUSING AND DATA MINING	L	T	P	J	C
		3	0	2	0	4
Pre-requisite	**Nil**	**Syllabus Revision**				
		V. 1.0				

Course Objectives:

1. To introduce the concept of Data Warehousing and Data Mining
2. To develop the knowledge for application of the mining algorithms for association, clustering
3. To explain the algorithms for mining data streams and the features of recommendation systems.

Expected Course Outcomes:

1. Interpret the contribution of data warehousing and data mining to the decision-support systems
2. Apply the link analysis and frequent item-set algorithms to identify the entities on the real world data
3. Apply the various classifications techniques to find the similarity between data items
4. Analyse the various data mining tasks and the principle algorithms for addressing the tasks
5. Evaluate and report the results of the recommended systems
6. Design the model to sample, filter and mine the Streaming data
7. Analyse the various data mining tasks for multimedia and complex data.

Student Learning Outcomes:	**2, 9, 12**

2. Having a clear understanding of the subject related concepts and of contemporary issues
9. Having problem solving ability- solving social issues and engineering problems
12. Having adaptive thinking and adaptability

Module 1	**DATA WAREHOUSE**	4 Hours

Introduction: Data Warehouse and OLAP Technology for Data Mining: Data Warehouse, Multidimensional Data Model, Data Warehouse Architecture, Data Warehouse Implementation, Further Development of Data Cube Technology, From Data Warehousing to Data Mining Data Cube Computation and Data Generalization: Efficient Methods for Data Cube Computation, Further Development of Data Cube and OLAP Technology, Attribute-Oriented Induction.

Module 2	**DATA PREPROCESSING**	4 Hours

Data, Types of Data, Attributes and Measurement, Types of Data Sets, Data Quality, Measurement and Data Collection Issues, Issues Related to Applications, Data pre-processing, Aggregation, Sampling, Dimensionality Reduction, Feature Subset Selection, Feature Creation, Discretization and Binarization, Variable Transformation, Similarity and Dissimilarity between Simple Attributes, Dissimilarities between Data Objects, Similarities between Data Objects.

Module 3	**ASSOCIATION ANALYSIS: CONCEPTS AND ALGORITHMS**	7 Hours

Frequent Itemset Generation, The Apriori Principle, Apriori Algorithm- Rule Generation- Candidate Generation and Pruning, Support Counting, Computational Complexity, Confidence-Based Pruning, Compact Representation of Frequent Itemsets, Maximal and Closed Frequent Itemsets, Alternative Methods for Generating Frequent Itemsets, FP-Growth Algorithm, FP-Tree Representation, Evaluation of Association Patterns, Handling Categorical Attributes, Handling Continuous Attributes, Discretization-Based Methods, Statistics-Based Methods, Non-discretization Methods, Sequential Pattern Discovery.

Module 4	**CLASSIFICATION AND PREDICTION**	7 Hours

Classification - issues regarding classification and prediction -Decision Tree Induction-Bayesian classification – Support Vector Machines, Rule-Based Classification- Associative Classification Prediction, Rationale for Ensemble Method, Methods for Constructing an Ensemble Classifier, Bias-Variance Decomposition, Bagging, Boosting, Random Forests, Empirical Comparison among Ensemble Methods

Module 5	**CLUSTER ANALYSIS AND OUTLIER ANALYSIS**	7 Hours
Types of Data in cluster analysis, - Major clustering methods- The k-Means Method, Agglomerative Hierarchical Clustering, Cluster Evaluation, Outlier Analysis- Distance-Based Outlier Detection-Density-Based Local Outlier Detection		
Module 6	**MINING OF STREAM DATA**	7 Hours
Mining Streams, Time Series and Sequence Data: Mining Data Streams, Mining Time-Series Data, Mining Sequence Patterns in Transactional Databases, Mining Sequence Patterns in Biological Data, Graph Mining, Social Network Analysis and Multi-relational Data Mining		
Module 7	**MULTIMEDIA AND COMPLEX DATA MINING**	7 Hours
Mining Object, Spatial, Multimedia, Text and Web Data: Multidimensional Analysis and Descriptive Mining of Complex Data Objects, Spatial Data Mining, Multimedia Data Mining, Text Mining, Mining the World Wide Web.		
Module 8	**RECENT TRENDS**	2 Hours
	Total Hours:	45 Hours

TEXT BOOKS:

1. Bhatia, Parteek, "Data mining and data warehousing: principles and practical techniques". Cambridge University Press, Ist Edition, 2019.
2. Karaa, Wahiba Ben Abdessalem, and Nilanjan Dey. *Mining multimedia documents*. CRC Press, 2017.

REFERENCE BOOKS:

1. Igual, Laura, and Santi Seguí. "Introduction to Data Science." In Introduction to Data Science, Springer, Cham, 2017.
2. Gupta, Gopal K. Introduction to data mining with case studies. PHI Learning Pvt. Ltd., 2014.
3. M. Kantardzic, "Data Mining: Concepts, Models, Methods, and Algorithms", 2nd edition, Wiley-IEEE Press, 2011.

Mode of Evaluation: CAT / Assignment / Quiz / FAT / Project / Seminar

List of Experiments

1.	Build Data Warehouse and Explore WEKA	3 hours
2.	Introduction to exploratory data analysis using R	3 hours
3.	Demonstrate the Descriptive Statistics for a sample data like mean, median, variance and correlation etc.,	3 hours
4.	Demonstrate Missing value analysis and different plots using sample data.	3 hours
5.	Demonstration of apriori algorithm on various data sets with varying confidence (%) and support (%).	3 hours
6.	Demo on Classification Techniques using sample data Decision Tree, ID3 or CART.	3 hours
7.	Demonstration of Clustering Techniques K-Mean and Hierarchical.	3 hours
8.	Demo on Classification Technique using KNN.	3 hours

9.	Demonstration on Document Similarity Techniques and measurements.	3 hours
10.	Demo on Classification Technique for multimedia data	3 hours
Mode of evaluation: Project/Activity		
Recommended by Board of Studies	Date: 11-02-2021	
Approved by Academic Council	No.61	Date: 18-02-2021