Course Description:

This course introduces advanced aspects of data warehousing and data mining, encompassing the principles, research results and commercial application of the current technologies.

Course Objective:

The main objective of this course is to provide knowledge of different data mining techniques and data warehousing.

Units and Unit Content

1. Introduction to Data Warehousing
teaching hours: 5 hrs
Lifecycle of data, Types of data, Data warehouse and data warehousing , Differences between operational database and data warehouse, A multidimensional data model, OLAP operation in multidimensional data model, Conceptual modeling of data warehouse, Architecture of data warehouse, Data warehouse implementation, Data marts, Components of data warehouse, Need for data warehousing ,Trends in data warehousing

2. Introduction to Data Mining
teaching hours: 2 hrs
Motivation for data mining, Introduction to data mining system, Data mining functionalities, KDD, Data object and attribute types, Statistical description of data, Issues and Applications

3. Data Preprocessing
teaching hours: 3 hrs
Data cleaning, Data integration and transformation, Data reduction, Data discretization and Concept Hierarchy Generation, Data mining primitives

4. Data Cube Technology
teaching hours: 4 hrs
Efficient method for data cube computation, Cube materialization (Introduction to Full cube, Iceberg cube, Closed cube, Shell cube), General strategies for cube computation, Attribute oriented induction for data characterization, Mining class comparison, Discriminating between
different classes

5. Mining Frequent Patterns
teaching hours: 6 hrs
Frequent patterns, Market basket analysis, Frequent itemsets, closed itemsets, association rules, Types of association rule (Single dimensional, multidimensional, multilevel, quantitative), Finding frequent itemset (Apriori algorithm, FP growth), Generating association rules from frequent itemset, Limitation and improving Apriori, From Association Mining to Correlation Analysis, Lift

6. Classification and Prediction (teaching hours: 10 hrs)
Definition (Classification, Prediction), Learning and testing of classification, Classification by decision tree induction, ID3 as attribute selection algorithm, Bayesian classification, Laplace smoothing, Classification by backpropagation, Rule based classifier (Decision tree to rules, rule coverage and accuracy, efficient of rule simplification), Support vector machine, Evaluating accuracy (precision, recall, f-measure), Issues in classification, Overfitting and underfitting, K-fold cross validation, Comparing two classifier (McNemar's test)

7. Cluster Analysis
teaching hours: 8 hrs
Types of data in cluster analysis, Similarity and dissimilarity between objects, Clustering techniques: - Partitioning (k-means, k-means++, Mini-Batch k-means, k-medoids), Hierarchical (Agglomerative and Divisive), Density based (DBSCAN), Outlier analysis

8. Graph Mining and Social Network Analysis
teaching hours: 5 hrs
Graph mining, Why graph mining, Graph mining algorithm (Beam search, Inductive logic programming), Social network analysis, Link mining, Friends of friends, Degree assortativity, Signed network (Theory of structured balance, Theory of status, Conflict between the theory of balance and status), Trust in a network (Atomic propagation, Propagation of distrust, Iterative propagation), Predicting positive and negative links

9. Mining Spatial, Multimedia, Text and Web Data
teaching hours: 2 hrs
Spatial data mining, Spatial data cube, Mining spatial association, Multimedia data mining, Similarity search in multimedia data, Mining association in multimedia data, An introduction to text mining, natural language processing and information extraction, Web mining (Web content mining, Web structure mining, Web usage mining)

Lab and Practical works

Laboratory Works:

The laboratory should contain all the features mentioned in a course, which should include data preprocessing and cleaning, implementing classification, clustering, association algorithms in any programming language, and data visualization through data mining tools.