ANDHRA UNIVERSITY
TRANS-DISCIPLINARY RESEARCH HUB

DATA WAREHOUSING AND MINING

UNIT I

Introduction: Fundamentals of data mining, Data Mining Functionalities, Classification of Data Mining systems, Data Mining Task Primitives, Integration of a Data Mining System with a Database or a Data Warehouse System, Major issues in Data Mining.

Data Preprocessing: Need for Preprocessing the Data, Data Cleaning, Data Integration and Transformation, Data Reduction, Discretization and Concept Hierarchy Generation.

UNIT II

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.

UNIT III

Mining Frequent Patterns, Associations and Correlations: Basic Concepts, Efficient and Scalable Frequent Itemset Mining Methods, Mining various kinds of Association Rules, From Association Mining to Correlation Analysis, Constraint-Based Association Mining **- Classification and Prediction:** Issues Regarding Classification and Prediction, Classification by Decision Tree Induction, Bayesian Classification, Rule-Based Classification, Classification by Backpropagation, Support Vector Machines, Associative Classification, Lazy Learners, Other Classification Methods, Prediction, Accuracy and Error measures, Evaluating the accuracy of a Classifier or a Predictor, Ensemble Methods

UNIT IV

Cluster Analysis Introduction : Types of Data in Cluster Analysis, A Categorization of Major Clustering Methods, Partitioning Methods, Hierarchical Methods, Density-Based Methods, Grid- Based Methods, Model-Based Clustering Methods, Clustering High-Dimensional Data, Constraint-Based Cluster Analysis, Outlier Analysis - 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 Multirelational Data Mining:

UNIT V

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. - **Applications and Trends in Data Mining:** Data Mining Applications, Data Mining System Products and Research Prototypes, Additional Themeson Data Mining and Social Impacts of Data Mining.

TEXT BOOKS:

1. Data Mining – Concepts and Techniques - Jiawei Han & Micheline Kamber, MorganKaufmann Publishers, 2nd Edition, 2006.
2. Introduction to Data Mining – Pang-Ning Tan, Michael Steinbach and Vipin Kumar, Pearsoneducation.

REFERENCES:

1. Data Warehousing in the Real World – Sam Aanhory & Dennis Murray Pearson Edn Asia.
2. Data Warehousing Fundamentals – Paulraj Ponnaiah Wiley student Edition
3. The Data Warehouse Life cycle Tool kit – Ralph Kimball Wiley student edition
4. Building the Data Warehouse By William H Inmon, John Wiley & Sons Inc, 2005.
5. Data Mining Introductory and advanced topics –Margaret H Dunham, Pearson education
6. Data Mining Techniques – Arun K Pujari, University Press.

ANDHRA UNIVERSITY

TRANS-DISCIPLINARY RESEARCH HUB

MODEL PAPER

DATA WAREHOUSING AND MINING

Answer any FIVE questions

All questions carry equal marks

1.a) Describe the steps involved in the process of knowledge discovery.

b) Explain the challenges to data mining regarding data mining methodology and user interaction issues. [10+10]

Use a flowchart to summarize the following procedures for attribute subset selection:

i) step-wise forward selection

ii) step-wise backward selection

2.Consider the following data for age and body fat, answer the following:

Age	23	27	39	41	47	49	50	52	54	56	57	58	58	60	61
%fat	26.5	17.8	31.4	25.9	27.4	27.2	31.2	34.6	42.5	33.4	30.2	34.1	32.9	41.2	35.7

i) Normalize the two attributes based on z-score normalization

ii) Calculate the correlation coefficient. Are these two attributes positively or negatively correlated? Compute their variance. [10+10]

3.a) Describe the situations where the query-driven approach is preferable to the update-driven approach.

b) Propose an algorithm that computes closed iceberg cubes efficiently. [10+10]

4.a) Discuss attribute oriented induction for class comparisons.

b) Differentiate between closed frequent itemset and maximal frequent itemset.How

c) to improve the efficiency of Apriori algorithm? [7+7+6]

5.a) Find the frequent itemsets in the following transactional database using a pattern-growth approach.

TID	List of Item IDs
T1	I1,I4,I5,I7,I8
T2	I4,I1,I9,I10,I2,I5
T3	I1,I2,I3,I4,I5,I10
T4	I2,I4,I5,I1,I7,I8
T5	I3,I5,I8,I9
T6	I1,I4,I5,I8,I7
T7	I2,I10,I9,I4
T8	I1,I2,I7,I9
T9	I2,I3,I4,I5,I9

b) Write about correlation analysis using chi-square measure. [10+10]

6.a) What is the need of decision tree pruning? Give an example.

b) What characteristics of neural networks make them a good classifier?
c) Write about bagging approach. [7+7+6]

7.a) Discuss the merits and demerits of hierarchical methods for clustering.
b) What is the role of graph mining in social network analysis?
c) Give the applications of outlier analysis. [7+7+6]

8.a) Explain HITS algorithm and compare it with PageRank algorithm.
b) Apply data mining techniques to spatial databases. [10+10]