CSA4201 Data Mining and Data Warehousing
Credits: 4C

Learning Outcomes	Suggested Pedagogical Processes
Students will be well aware of structure of Data Warehouse and the ETL process.	Power point presentation for structure of data warehouse and for ETL process the case study should be presented.
Students will get know about different pre processing methodologies.	With the help of various free tools students will be asked to smooth the data. For the same sake different data sets will be used.
Students should able to do some basic descriptive and predictive data mining. And able to compare and contrast different concepts.	Power point presentation to describe the working of the technique and show R code to understand the working of the algorithms.
Students should able to evaluate supervised and unsupervised models based on the accuracy.	Statistical approaches to find the accuracy, various case studies will be presented to evaluate the accuracy.
Students should able to analyse the given problem and using their skills able to solve the practical problem.	Different tools and various readymade datasets will be presented, so that students could able to find out some trends in the data. Classroom discussions and learning by communicating interest will be applied.

Unit No.	Title of Unit and Contents
I	**Introduction to Data Mining** 1.1 Definition of Data Mining and Data Warehousing 1.2 DM versus Knowledge 1.3 Discovery in Databases 1.4 Data to be mined 1.5 Basic mining techniques 1.6 Data Mining Issues 1.7 Data Mining Metrics 1.8 Social Implications of Data Mining 1.9 Overview of Applications of Data Mining
II	**Data Pre-processing** 2.1 Data Processing prerequisites 2.2 Attributes and Data types 2.3 Statistical descriptions of data 2.4 Distance and similarity measures 2.5 Need for Preprocessing 2.6 Handling Missing data 2.7 Data Cleaning 2.8 Data Integration 2.9 Data Reduction 2.10 Data Transformation and Data Discretization
III	**Introduction to Data Warehousing** 3.1 Architecture of DW 3.2 OLAP and Data Cubes 3.3 Dimensional Data Modeling-star, snowflake schemas

	3.4 Concept of data mart
IV	**Association Rule Mining** 4.1 Market Basket analysis 4.2 Frequent item-sets 4.3 Association rule mining: Apriori algorithm, FP growth algorithm, Sampling Algorithms
V	**Classification & Prediction** 5.1 Definition of classification 5.2 Model construction 5.3 Model Usage 5.4 Choosing algorithm 5.5 K-nearest neighbor algorithm 5.6 Decision tree Induction 5.7 Information gain 5.8 gain ratio 5.9 gini index 5.10 Bayesian Classification 5.11 Bayes Theorem 5.12 Naïve Bayes classifier 5.13 Measuring performance of classifiers 5.14 Precision 5.15 Recall 5.16 F-measure 5.17 confusion matrix 5.18 cross-validation 5.19 Bootstrap 5.20 Linear Regression 5.21 Non-linear Regression 5.22 Logistic Regression
VI	**Clustering** 6.1 Definitions 6.2 Partitioning methods 6.3 Hierarchical clustering 6.4 Density Based methods
VII	**Data Mining Tool** 7.1 Weka 7.2 Performance measures TP, FP, ROC 7.3 Baseline algorithms zeroR, oneR

Learning Resources

1. Tom Mitchell, Machine Learning, McGraw Hill, 1997
2. R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, Second edition, 2011
3. Jiawei Han, Micheline Kamber, Jian Pei, Data Mining: Concepts and Techniques, ISBN:9789380931913, Elsevier Morgan Kaumann Publishers, 3rd Ed., 2012
4. Margaret H. Dunham, S. Sridhar, Data Mining - Introductory and Advanced Topics, Pearson Education, 2012
5. George Marak, Modern Data warehousing and mining and visualization, Pearson Publication