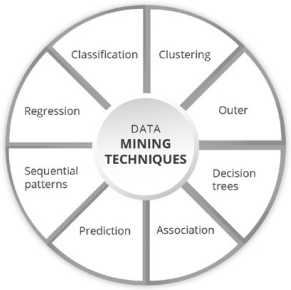

Source: https://alternative-spaces.com/blog/8-data-mining-techniques-you-must-learn-to-succeed-in-business/

22CS306 DATA MINING TECHNIQUES

Hours Per Week :

L	T	P	C
2	0	2	3

PREREQUISITE KNOWLEDGE: Probability and statistics, Python programming.

COURSE DESCRIPTION AND OBJECTIVES:

This course introduces the basic concepts, principles, methods, implementation techniques, and applications of data mining, with a focus on three major data mining functions: (1) Association rule mining (2) Classification and (3) cluster Analysis. It also focuses on issues relating to the feasibility, usefulness, effectiveness and scalability of techniques for the discovery of patterns hidden in large data sets.

MODULE-1

UNIT-1 **8L+0T+8P=16 Hours**

INTRODUCTION

What is Data Mining? Why data mining?; What kinds of data can be mined?;What kinds of patterns can be mined?; Which technologies are used?; What kinds of applications are targeted? Major issues in data mining; Data objects and attribute types; Basic statistical descriptions of data, Data matrix versus dissimilarity matrix.

Data Pre-processing: Overview - data quality, major tasks in data preprocessing; Data cleaning - missing values, noisy data; Data Integration - entity identification problem, redundancy and correlation analysis tuple duplication; Data value conflict detection and resolution; Data reduction - PCA, attribute subset selection, regression and log linear models; Histogram; Data transformation - data transformation by normalization; Discretization by binning; Discussion on ethical considerations related to collection, analysis, and use of data. Introduction to privacy-preserving in datamining.

UNIT-2 **8L+0T+8P=16 Hours**

ASSOCIATION ANALYSIS

Market basket analysis; Frequent Item sets; Closed item sets and association rules; Frequent Item set Mining Methods-apriori algorithm, generating association rules, improving apriori, FP growth method, vertical format method; Which patterns are interesting? Pattern evaluation method; Pattern Mining in multilevel multidimensional space, Pattern Mining in Multilevel, Multidimensional Space.

PRACTICES:

- Apply the following data pre-processing techniques on dataset (download from n UCI/ Kaggle/
- NCBI data repository) to illustrate the need of the pre-processing in data mining
  a) Data Cleaning
  b) Data Normalization
  c) Data Discretization
  d) Computation of correlation coefficient to analyze the data behavior
  e) Dimensionality reduction using PCA and Wavelets
- Construct Heat Map Table to understand the Correlation among the attributes in a given dataset.
- Extract the interesting association rules from a given dataset using A priori algorithm.
- Extract the interesting association rules from a given dataset using Frequent Pattern growth algorithm.

MODULE-2

UNIT-1 **8L+0T+8P=16 Hours**

CLASSIFICATION

What is classification?, General approach to classification, Decision tree induction - attribute selection measures; Tree pruning; Bayes Classification methods - Bayes theorem; Naïve Bayesian classification; Classification by back propagation - a multilayer feed forward neural network; Defining a network topology; Back propagation; K nearest neighbor classifier; Support vector machine, Linearly separable and inseparable cases, Model evaluation and selection; Techniques to improve classification accuracy.

UNIT-2 **8L+0T+8P=16 Hours**

CLUSTER ANALYSIS

Partition methods - K means and K medoid; Hierarchical methods; Agglomerative and divisive method; Density based methods - DBSCAN; Optics; Grid based methods-STING; Cluster evaluation methods; Clustering high dimensional data; Problems, Challenges and major methodologies;

Global and Local Impact Analysis: Health care and Environmental Sustainability (Ex: Weather forecasting)

PRACTICES:

- Apply the following classifiers on a given dataset and analyze their performance.
  a) J48 and visualize the decision tree.
  b) Naive Bayes.
  c) Support Vector Machine.
  d) Multi-Layer Perceptron.
  e) K-Nearest Neighbor.
- Illustrate the performance of Ensemble Classification algorithms such as Bagging and Boosting Methods.
- Evaluate the performance of partitioning clustering algorithms on a given dataset.
- Evaluate the performance of hierarchical clustering algorithms on a given dataset.

COURSE OUTCOMES:

Upon successful completion of this course, students will have the ability to:

CO No.	Course Outcomes	Blooms Level	Module No.	Mapping with POs
1	Investigate various patterns that can be extracted from different types of data.	Analyze	1,2	1, 2,
2	Apply various pre-processing techniques and classification algorithms on different domains of data.	Apply	1,2	1, 2, 5, 6
3	Build decision making systems using data mining algorithms for a given real time data set.	Apply Create	1,2	3, 5, 8
4	Construct models using modern tools such as WEKA, R and Python etc.	Apply Create	1,2	1, 2,5,9

TEXT BOOKS:

1. Jiawei Han, Micheline Kamber and Jian Pei, “Data mining Concepts and Techniques”, 3rd Edition, Morgan Kaufmann. 2012.
2. Pang-Ning Tan, Michael Steinbach and Vipin Kumar, “Introduction to Data Mining”, 2nd Edition, Pearson, 2018.

REFERENCE BOOKS:

1. Jure Leskovec, Anand Raja raman and Jeffrey D Ullman, “Mining of Massive Datasets”, 5th Edition, Stanford University, 2014.
2. GK Gupta, Introduction to Data Mining with Case Studies, Prentice Hall. 3rd Edition,2014.
3. Margaret H Dunham, “Data Mining: Introductory and Advanced Topics”, PEA, 2008.

SKILLS:

- ✓ *Handle various types of Data and able to explore the characteristics of data*
- ✓ *Perform various Data Visualisation tasks over the data and present the data with ease of access*
- ✓ *Perform descriptive and predictive mining tasks over the data to carry out decision making.*