Nirma University
Institute of Technology, School of Technology
M. Tech. Computer Science and Engineering (Data Science)
Semester – II

L	T	P	C
3	0	2	4

Course Code	3CS42D105
Course Name	Data Mining and Visualization

Course Learning Outcomes (CLOs):

At the end of the course, students will be able to

1. identify a number of common data domains and corresponding analysis tasks, including multivariate data, networks, text and cartography
2. comprehend the key processes of data mining, data warehousing and knowledge discovery process
3. implement data mining techniques to solve problems in other disciplines in a mathematical way
4. exercise building and evaluating visualization systems

Syllabus:

Teaching Hours

Unit I **10**

Data Understanding: types of data, information and uncertainty, classes and attributes, interactions among attributes, relative distributions, summary statistics.
Data Visualization: using different tools - refine data and create, edit, alter, and display their visualizations (x-y graph, bar chart, pie chart, cube etc)
Data Quality: inaccurate data, sparse data, missing data, insufficient data, imbalanced data
Social Challenges: data ownership, data security, ethics and privacy

Unit II **15**

Data Reduction and Feature Enhancement: standardizing data, sampling data, using principal components to eliminate attributes, limitations and pitfalls of principal component analysis (PCA), curse of dimensionality
Clustering: dissimilarity and scatter, categorization, k-means clustering, hierarchical clustering, distance measures, shape of clusters, determining the number of clusters, evaluating clusters
Association Analysis: association rule learning, the Apriori algorithm, FP-Growth, market basket analysis

Unit III **15**

Machine Learning Algorithms for Data Mining: Regression, review of linear regression, assumptions underlying linear regression. Classification, supervised categorization, linear classifiers, logistic regression, regression trees, classification trees, Bayes' Theorem, Na ̈ive Bayes, support vector machines, confusion matrices, receiver operating characteristic (ROC) curves, precision and recall, lift curves, cost curves
Model Selection and Validation: training error and optimism, the Bayes error rate,

inductive bias, the bias-variance trade off, overfitting, Occam's Razor, minimum description length (MDL), sampling bias, the validation set approach, leave-one-out cross-validation, k-fold cross-validation, boot strapping, jack knifing, data snooping

Ensemble Learning: bootstrap aggregating (bagging), boosting, stacking/blending, random subspaces, random forests.

Unit IV **5**

Recommender Systems, Reinforcement Learning, Active Learning, Semi-supervised Learning, Transfer Learning, Deep Learning, Data Stream Mining

Self-Study:

The self-study contents will be declared at the commencement of semester. Around 10% of the questions will be asked from self-study contents.

Laboratory Work:

Laboratory work will be based on above syllabus with minimum 5 experiments to be incorporated.

Suggested Readings^:

1. Jiahei Han & Micheline Kamber, Data Mining Concepts and Techniques, Morgan Kaufmann
2. Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Introduction to Data Mining, Pearson
3. Wes McKinney, Python for Data Analysis, Oreilly
4. S. Nagabhushana, Data Warehousing OLAP and Data Mining, New Age publishers

L=Lecture, T=Tutorial, P=Practical, C=Credit

^this is not an exhaustive list