3. Run Apriori algorithm to find frequent itemsets and association rules on 2 real datasets and use appropriate evaluation measures to compute correctness of obtained patterns a) Use minimum support as 50% and minimum confidence as 75% b) Use minimum support as 60% and minimum confidence as 60 % I.
4. Use Naive bayes, K-nearest, and Decision tree classification algorithms and build classifiers on any two datasets. Divide the data set into training and test set. Compare the accuracy of the different classifiers under the following situations: a) Training set = 75% Test set = 25% b) Training set = 66.6% (2/3rd of total), Test set = 33.3% II. Training set is chosen by i) hold out method ii) Random subsampling iii) Cross-Validation. Compare the accuracy of the classifiers obtained. Data is scaled to standard format.
5. Use Simple K-means algorithm for clustering on any dataset. Compare the performance of clusters by changing the parameters involved in the algorithm. Plot MSE computed after each iteration using a line plot for any set of parameters.

Project: Students should be promoted to take up one project on any UCI/kaggle/data.gov.in or a dataset verified by the teacher. Preprocessing steps and at least one data mining technique should be shown on the selected dataset. This will allow the students to have a practical knowledge of how to apply the various skills learnt in the subject for a single problem/project.

Note: Examination scheme and mode shall be as prescribed by the Examination Branch, University of Delhi, from time to time.

DSE-A4/DSE: DATA MINING-II

CREDIT DISTRIBUTION, ELIGIBILITY AND PRE-REQUISITES OF THE COURSE

Course title & Code	Credits	Credit distribution of the course			Eligibility criteria	Pre-requisite of the course (if any)
		Lecture	Tutorial	Practical/ Practice		

Data Mining - II	**4**	**3**	**0**	**1**	Pass in Class XII	Data Mining-I

Course Objectives

The course introduces the students to the supervised and unsupervised learning techniques. Students will learn about the importance of ensemble methods, cluster analysis, anomaly detection and their applicability in mining patterns in real applications. At the end students will be exposed to two advanced topics: text mining and time-series mining. Students will use the learned topics in solving real applications using open-source software.

Learning outcomes

On successful completion of the course, students will be able to:

● Differentiate between partition-based, density-based and hierarchical clustering

● Build ensemble models to improve predictive performance of the classifier

● Identify anomalies and outliers using supervised and unsupervised techniques

● Analyze time-series data and extract patterns from the stamped data

● Mine textual data and do topic modelling

Syllabus

Unit 1 **(9 hours)**

Clustering:

Partitioning Methods, Hierarchical Methods, Density-Based Methods, Comparison of different methods

Unit 2 **(8 hours)**

Ensemble Methods:

Need of ensemble, Random Forests, Bagging and Boosting

Unit 3 **(10 hours)**

Anomaly Detection:

Outliers and Outlier Analysis, Outlier Detection Methods, Statistical Approaches, Proximity-based and density-based outlier detection, Clustering-based approaches

Unit 4 **(8 hours)**

Mining Text Data:

Document Preparation and Similarity, Clustering Methods for Text, Topic Modeling

Unit 5 **(10 hours)**

Stream Mining:

Time series basics, Date Ranges, Frequencies, and Shifting, Resampling and moving windows functions, Decay function, Clustering stamped data: STREAM and CluStream

Essential/recommended readings

1. Tan P.N., Steinbach M, Karpatne A. and Kumar V. Introduction to Data Mining, 2nd edition, Pearson, 2019.

2. Zaki M. J. and Meira J. Jr. Data Mining and Machine Learning: Fundamental Concepts and Algorithms, 2nd edition, Cambridge University Press, 2020.

3. Aggarwal C. C. Data Mining: The Textbook, Springer, 2015.

Additional References

1. Han J. Kamber M. and Pei J. Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, 2011.

2. Dunham M. Data Mining: Introductory and Advanced Topics, Pearson, 2006

Suggested Practicals List (If any): (30 Hours)

Practical exercise such as

1. Apply Partitioning Methods, Hierarchical Methods, Density-Based Methods for clustering on a data set and compare the performance of the obtained results using different metrics
2. Create an ensemble using Random Forest and show the impact of bagging and boosting on the performance
3. Apply different outlier-detection methods on a noisy dataset and compare their effectiveness in terms of outliers reported
4. Compute similarity between two documents after required document preparation
5. Considering a time-stamped data (sales data/weather data), compare the aggregate values visually using different moving windows function
6. Write a program to find the latent topics in a document using any topic modeling method and display top 5 terms that contribute to each topic along with their strength. Also, visualize the distribution of terms contributing to the topics.

Project: Students should be promoted to take up one project covering at least one unit of the syllabus on any UCI/kaggle/data.gov.in or a dataset verified by the teacher. This will allow the students to have a practical knowledge of how to apply the various skills learnt in the subject for a single problem/project.

GE4b/DSE: INTRODUCTION TO WEB PROGRAMMING

Credit distribution, Eligibility and Pre-requisites of the Course

Course title & Code	Credits	Credit distribution of the course			Eligibility criteria	Pre-requisite of the course (if any)
		Lecture	Tutorial	Practical/ Practice		
Introduction to web programming	4	3	0	1	Pass in Class XII	NIL

Course Objectives