BSc(H) Computer Science

DISCIPLINE SPECIFIC Elective- Data Mining-II (Guidelines)
Sem V (July 2024 Onwards)

Sr. No.	Units	Chapter	Reference	No. of Hours
1	**Unit 1: Clustering**: Partitioning methods, hierarchical methods, density-based methods, comparison of different methods	5.2.1, 5.2.5, 5.3 (5.3.1, 5.3.2, 5.3.4, 5.3.5, 5.3.6), 5.4,5.5.7	[1]	9
2	**Unit 2: Ensemble Methods**: Need of ensemble, random forests, bagging and boosting	6.10, 6.11 (introduction, 6.11.2)	[1]	8
3	**Unit 3: Anomaly Detection**: Outliers and outlier analysis, outlier detection methods, statistical approaches, proximity-based and density-based outlier detection, clustering-based approaches	9.1, 9.2, 9.3 (9.3.1, 9.3.2, 9.3.5), 9.4, 9.5	[1]	10
4	**Unit 4: Mining Text Data**: Document preparation and similarity, clustering methods for text, topic modeling	13.1, 13.2, 13.2.1, 13.3, 13.3.1 (excluding its subsection), 13.3.3, 13.4 (Upto Page 441)	[2]	8
5	**Unit 5: Stream Mining**: Time series basics, date ranges, frequencies, shifting, resampling and moving windows functions, decay function, clustering stamped data: STREAM and CluStream	11.1, 11.2, 11.3,11.6,11.7 2.2.2.4, 2.2.2.5, 2.4.1.1, 12.4.1-12.4.2	[3] [2]	10

Text Book:

1. Tan P.N., Steinbach M, Karpatne A. and Kumar V. Introduction to Data Mining, Second edition, Sixth Impression, Pearson, 2023.
2. Aggarwal C. C. *Data Mining: The Textbook*, Springer, 2015
3. McKinney W. *Python for Data Analysis: Data Wrangling with Pandas*, *NumPy and IPython*. 2nd edition. O'Reilly Media, 2018.

Additional References:

1. Han J., Kamber M. and Pei J. *Data Mining: Concepts and Techniques*, 3rd edition, 2011, Morgan Kaufmann Publishers.

2. Zaki M. J. and Meira J. Jr. *Data Mining and Machine Learning: Fundamental Concepts and Algorithms*, 2nd edition, Cambridge University Press, 2020.
3. Insight into Data mining: Theory and Practice, Soman K. P., Diwakar Shyam, Ajay V., PHI 2006

For practicals, datasets may be downloaded from :

1. https://archive.ics.uci.edu/datasets
2. https://www.kaggle.com/datasets?fileType=csv
3. https://data.gov.in/
4. https://ieee-dataport.org/datasets
5. Time Series Datasets (kaggle.com)

Suggested Practical Exercises

1. Perform partitioning, hierarchical, and density-based clustering algorithms on a downloaded dataset and evaluate the cluster quality by changing the algorithm's parameters.
2. Perform the following text mining preprocessing steps on a text document:
    a. Stop Word Removal
    b. Stemming
    c. Removal of punctuation marks
    d. Compute the inverse document frequency of the words in the document
3. Use the Decision Tree classification algorithm to construct a classifier on two datasets. Evaluate the classifier's performance by dividing the dataset into a training set (75%) and a test set (25%). Compare the performance with that of:
    a. Bagging ensemble consisting of 3,5,7,9 Decision tree classifiers
    b. Adaboost ensemble consisting of 3,5,7,9 Decision tree classifiers
4. Download a dataset and check whether outliers are present in the dataset. Use different methods of outlier detection and compare their performance.
5. Perform CluStream algorithm on any time series data from Kaggle and compare its output with that of K-means clustering. Evaluate the cluster quality by changing the algorithm's parameters.

Project: *Students should be promoted to take up one project on a dataset downloaded from any of the websites given above and the dataset verified by the teacher. Apply at least two data mining concepts on the selected dataset.*

Prepared by:

1. Dr Anamika Gupta (Shaheed Sukhdev College of Business Studies)
2. Dr Manju Bhardwaj (Maitreyi College)
3. Dr Sarabjeet Kaur (Indraprastha College For Women)
4. Prof. Sharanjit Kaur (Acharya Narendra Dev College)