B.A. with Computer Science as Major discipline

Undergraduate Programme of study with Computer Science discipline as one of the two Core Disciplines

DISCIPLINE SPECIFIC CORE COURSE- Data Mining-I (Guidelines)
Sem III (August 2023)

Sr. No.	Units	Chapter	No. of Hours
1	**Unit 1: Introduction to Data Mining:** Motivation and Challenges for data mining, Types of data mining tasks, Applications of data mining, Data measurements, Data quality, Supervised vs. unsupervised techniques	1.1-1.4, 2.1-2.2	8
2	**Unit 2: Data Pre-Processing:** Data aggregation, sampling, dimensionality reduction, feature subset selection, feature creation, variable ransformation.	2.3.1, 2.3.2, 2.3.3 (introduction), 2.3.4 (introduction), 2.3.5 (introduction), 2.3.6 (Binarization and Discretization of Continuous attributes), 2.3.7, 2.4.2, 2.4.3 (*excluding properties)*	9
3	**Unit 3: Cluster Analysis:** Basic concepts of clustering, measure of similarity, types of clusters and clustering methods, K-means algorithm, measures for cluster validation, determine optimal number of clusters	7.1.1, 7.1.2, 7.1.3 (well-separated and Density-based) 7.2 (*upto Data in Euclidean Space*), 7.5.1, 7.5.5	11
4	**Unit 4: Association Rule Mining:** Transaction data-set, frequent itemset, support measure, rule generation, confidence of association rule, Apriori algorithm, Apriori principle	5 (*up to* 5.2.2)	8
5	**Unit 5: Classification:** Naive Bayes classifier, Nearest Neighbour classifier, decision tree, overfitting, confusion matrix, evaluation metrics and model evaluation.	3 (*up to* 3.3.3), 3.4 (introduction) 3.6, 4.3, 4.4	9

Text Book:

1. Tan P.N., Steinbach M, Karpatne A. and Kumar V. *Introduction to Data Mining,* Second edition, Pearson, 2021.

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. Aggarwal C. C. *Data Mining: The Textbook*, Springer, 2015
4. Insight into Data mining: Theory and Practice, Soman K. P., Diwakar Shyam, Ajay V., PHI 2006

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

Suggested Practical Exercises

1. Apply data cleaning techniques on any dataset (e,g, wine dataset). Techniques may include handling missing values, outliers, inconsistent values. A set of validation rules can be prepared based on the dataset and validations can be performed.

2. Apply data pre-processing techniques such as standardization/normalization, transformation, aggregation, discretization/binarization, sampling etc. on any dataset

3. Run Apriori algorithm to find frequent item sets 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 %

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:

I. 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 needs to be 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 using dataset downloaded from any of the websites given above and the 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.*

Prepared by:

Dr Anamika Gupta (Shaheed Sukhdev College of Business Studies)

Dr Manju Bhardwaj (Maitreyi College)

Dr Sarabjeet Kaur (Indraprastha College For Women)

Prof. Sharanjit Kaur (Acharya Narendra Dev College)