BSc Program Computer Science

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

DISCIPLINE SPECIFIC Elective- Data Mining and Knowledge Discovery (Guidelines) Sem V (July 2024 onwards)

Sr. No.	Units	Chapter	Reference	No. of Hours
1	**Unit 1 Introduction** Need for data mining, Data mining tasks, Applications of data mining, Measures of similarity and dissimilarity, Supervised vs. unsupervised techniques.	1.1-1.4, 2.4.2, 2.4.3 *(excluding properties)*	[1]	6
2	**Unit 2 Data collection and preparation** Measurement and data collection issues, Data aggregation, Sampling, Dimensionality reduction, Feature subset selection, Feature creation, Discretization and binarization, Variable transformation..	2.1,2.2, 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	[1]	8
3	**Unit 3 Clustering data (14 Hours)** Basic concepts of clustering, Partitioning Methods: K-means algorithm, Hierarchical Methods: Agglomerative Hierarchical Clustering, Density-Based Methods: DBSCAN Algorithm, Strengths and weaknesses of different methods, Cluster evaluation.	5.2 (5.2.1-*upto Data in Euclidean Space*, 5.2.5), 5.3 (5.3.1, 5.3.2-*Excluding Ward's and Centroid methods*, 5.3.6), 5.4, 5.5(5.5.1,5.5.5,5.5.7)	[1]	14
4	**Unit 4 Classification** Preliminaries, Naive Bayes classifier, Nearest Neighbour classifier, Decision tree, Artificial Neural Network, overfitting, Confusion matrix, Evaluation metrics and Model evaluation.	3 (*up to* 3.3.3), 3.4 (introduction) 3.6, 6.3, 6.4, 6.7 (introduction), 6.11(introduction, 6.11.2)	[1]	10
5	**Unit 5 Ensemble Methods** Need for ensembles, Random Forest, Concept of Bagging and Boosting in ensembles.	6.10 (Excluding 6.10.3)	[1]	7

Text Book:

1. Tan P.N., Steinbach M, Karpatne A. and Kumar V. Introduction to Data Mining, Second edition, Sixth Impression, Pearson, 2023.

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. Dunham M. *Data Mining: Introductory and Advanced Topics*, Pearson, 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. Apply data cleaning techniques on any dataset (e.g. Chronic Kidney Disease dataset from UCI repository). Techniques may include handling missing values, outliers and inconsistent values. Also, a set of validation rules may be specified for the particular dataset and validation checks performed.

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

3. Apply simple K-means algorithm for clustering any dataset. Compare the performance of clusters by varying the algorithm parameters. For a given set of parameters, plot a line graph depicting MSE obtained after each iteration.

4. Perform partitioning, hierarchical, and density-based clustering algorithms on a downloaded dataset and evaluate the cluster quality by changing the algorithm's parameters.

5. Use Naive bayes, K-nearest, and Decision tree classification algorithms to build classifiers on any two datasets. Pre-process the datasets using techniques specified in Q2. Compare the Accuracy, Precision, Recall and F1 measure reported for each dataset using the abovementioned classifiers under the following situations:
   i. Using Holdout method (Random sampling):
      a) Training set = 80% Test set = 20%
      b) Training set = 66.6% (2/3rd of total), Test set = 33.3%
   ii. Using Cross-Validation:
      a) 10-fold
      b) 5-fold

6. Use the Decision Tree classification algorithm to construct a classifier on two datasets. Evaluate the classifier's performance by performing ten-fold cross validation. Compare the performance with that of:
   i. Bagging ensemble consisting of 3, 5, 7, 9 Decision tree classifiers
   ii. Adaboost ensemble consisting of 3, 5, 7, 9 Decision tree classifiers

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:

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)