Guidelines Data Mining B.Sc. Physical Science Semester III/V (NEP UGCF 2022

Discipline Specific Elective Semester III (DSE/A3)

# (Effective from Academic Year 2025-26)

# Name of the Teacher: - Dr. Dharmendera Kumar Meena

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sr.**  **No.** | **Units** | **Chapter** | **Refer ence** | **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 | [1] | 7 |
| 2 | **Unit 2: Data Pre-processing:** Data aggregation, sampling, dimensionality reduction, feature subset selection, feature creation, variable transformation. | 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)* | [1] | 8 |
| 3 | **Unit 3: Cluster Analysis:** Basic concepts of clustering, measure of similarity, types of clusters and clustering methods, Distance-based method: K-means algorithm, measures for cluster validation, determine optimal number of clusters. Density-Based Method: DBSCAN Algorithm, Comparison of these two methods | 5.1.1, 5.1.2, 5.1.3 (well-separated  and Density-based), 5.2 (5.2.1- *upto Data in Euclidean Space*, 5.2.5), 5.4, 5.5 (5.5.1,5.5.5,5.5.7) | [1] | 11 |
| 4 | **Unit 4: Association Rule Mining:** Transaction data-set, frequent itemset, support measure, rule generation, confidence of association rule, Apriori principle, Apriori algorithm | 4 (*up to* 4.2.2), 4.3 (introduction,  4.3.1) | [1] | 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, 6.3, 6.4, 6.11 (introduction,  6.11.2) | [1] | 11 |

# 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.

1. Zaki M. J. and Meira J. Jr. *Data Mining and Machine Learning: Fundamental Concepts and Algorithms*, 2nd edition, Cambridge University Press, 2020.
2. Aggarwal C. C. *Data Mining: The Textbook*, Springer, 2015

# 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>

# Suggested Practical Exercises

1. Apply data cleaning techniques on any dataset (e.g., Paper Reviews dataset in UCI repository). Techniques may include handling missing values, outliers and 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
   1. Use minimum support as 50% and minimum confidence as 75%
   2. Use minimum support as 60% and minimum confidence as 60 %
4. 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:
5. Using Holdout method (Random sampling):
   1. Training set = 80% Test set = 20%
   2. Training set = 66.6% (2/3rd of total), Test set = 33.3%
6. Using Cross-Validation:
   1. 10-fold
   2. 5-fold
7. 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.
8. Perform density-based clustering algorithm on a downloaded dataset and evaluate the cluster quality by changing the algorithm's parameters.