Department of Computer Science

COURSES OFFERED BY DEPARTMENT OF COMPUTER SCIENCE

(Computer Science Courses for Undergraduate Programme of study with Computer Science discipline as one of the three Core Disciplines)

Semester 5

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		
DSE 03a : Data Mining for Knowledge Discovery	**4**	**3**	**0**	**1**	Pass in Class XII	**Knowledge of Python Programming**

Learning Objectives

This course aims to introduce supervised and unsupervised data mining techniques, and their applications to real-life datasets. The students will learn about data quality, how to pre-process a dataset to make it ready for the application of data mining algorithms. The course will primarily focus on two core data mining techniques: classification and clustering. Different algorithms under the ambit of these techniques will be discussed, along with their strengths and weaknesses, model evaluation, and result evaluation metrics. The importance of ensemble methods, random forests, and the use of bagging and boosting in ensembles will be explained. The students will be

encouraged to apply the aforementioned data mining concepts to real life problems using open-source software.

Learning outcomes

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

1. Pre-process the data for subsequent data mining tasks
2. Apply a suitable classification algorithm to train the classifier and evaluate its performance.
3. Apply appropriate clustering algorithm to cluster the data and evaluate clustering quality
4. Differentiate between partition-based, density-based and hierarchical clustering
5. Build ensemble models to improve predictive performance of a classifier
6. Apply appropriate data mining techniques to solve a real-life problem

SYLLABUS OF DSE 03a

Unit 1 Introduction (6 Hours)

Need for data mining, Data mining tasks, Applications of data mining, Measures of similarity and dissimilarity, Supervised vs. unsupervised techniques.

Unit 2 Data collection and preparation (8 Hours)

Measurement and data collection issues, Data aggregation, Sampling, Dimensionality reduction, Feature subset selection, Feature creation, Discretization and binarization, Variable transformation..

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.

Unit 4 Classification (10 Hours)

Preliminaries, Naive Bayes classifier, Nearest Neighbour classifier, Decision tree, Artificial Neural Network, overfitting, Confusion matrix, Evaluation metrics and Model evaluation.

Unit 5 Ensemble Methods (7 Hours)

Need for ensembles, Random Forest, Concept of Bagging and Boosting in ensembles.

Essential/recommended readings

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

Online references/material:

1. http://www.dcc.fc.up.pt/~ltorgo/DM1/dataPreProc.html
2. https://www.coursera.org/specializations/data-mining-foundations-practice

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

All topics covered in theory will be implemented using Python. The operations may be performed on the datasets loaded through scikit, seaborn libraries or can be downloaded from from Open Data Portal (https:// data.gov.in/, UCI repository http://archive.ics.uci.edu/ml/ **).**

Recommended Datasets for :

Classification: Abalone, Artificial Characters, Breast Cancer Wisconsin (Diagnostic)

Clustering: Grammatical Facial Expressions, HTRU2, Perfume data

Suggestive practicals include:

1. Apply data cleaning techniques on any dataset (eg. 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. Use Simple K-means algorithm for clustering on any dataset. Compare the performance of clusters by changing the parameters involved in the algorithm. Plot Mean Squared Error computed after each iteration using a line plot for any set of parameters

4. 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
5. Create an ensemble using Random Forest and show the impact of bagging and boosting on the performance
6. 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 should be chosen by i) hold out method ii) Random subsampling iii) Cross-Validation. Compare the accuracy of the classifiers obtained.
   III. Data should be scaled to the standard format.

Project

Students should be promoted to take up one project on any UCI/kaggle/data.gov.in or on 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 practical knowledge of how to apply the various skills learned in the subject to a single problem/project.

DISCIPLINE SPECIFIC Elective (DSE 03b): Web Design and Development

Semester 5

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		
DSE 03b Web Design and Development	**4**	**3**	**0**	**1**	Pass in Class XII	Knowledge of Structured Query Language (SQL)