DISCIPLINE SPECIFIC ELECTIVE COURSE – DSE 3A Data Mining

Course title &Code	Credits	Credit distribution of the course			Eligibility criteria	Pre- requisite of the course (if any)
		Lecture	Tutorial	Practical/ Practice		
Data Mining	4	3	0	1	Class XII Pass	NA

Learning Objectives

This course introduces data mining techniques and enables students to apply these techniques on real-life datasets. The course focuses on three main data mining techniques: Classification, Clustering and Association Rule Mining tasks.

Learning Outcomes

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

1. Pre-process the data, and perform cleaning and transformation
2. Apply suitable classification algorithm to train the classifier and evaluate itsperformance.
3. Apply appropriate clustering algorithm to cluster data and evaluate clusteringquality.
4. Use association rule mining algorithms and generate frequent item-sets andassociation rules.

SYLLABUS OF DSE-3A

Unit I Introduction to Data Mining (9 Hours)

Applications of data mining, data mining tasks, motivation and challenges, types of data attributes and measurements, data quality.

Data Pre-processing - aggregation, sampling, dimensionality reduction, Feature Subset Selection, Feature Creation, Discretization and Binarization, VariableTransformation.

Unit 2 Classification and Model Evaluation (12 Hours)

Basic Concepts, Decision Tree Classifier: Decision tree algorithm, attributeselection measures, Nearest Neighbour Classifier, Bayes Theorem and Naive Bayes Classifier, Holdout Method, Random Sub Sampling, Cross-Validation, evaluation metrics, confusion matrix.

Unit 3 Association rule mining (12 Hours)

Transaction data-set, Frequent Itemset, Support measure, Apriori Principle, Apriori Algorithm, Computational Complexity, Rule Generation, Confidence of association rule.

Unit 4 Cluster Analysis (12 Hours)

Basic Concepts, Different Types of Clustering Methods, Different Types of Clusters, K-means: The Basic K-means Algorithm, Strengths and Weaknesses of K-means algorithm, Agglomerative Hierarchical Clustering: Basic Algorithm, Proximity between clusters.

Essential Readings

1. Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, Vipin Kumar,Introduction to Data Mining, 2nd Edition, Pearson, 2021.
2. Jiawei Han, Micheline Kamber, Pei Jian, Data Mining: Concepts andTechniques. 3rd edition. Morgan Kaufmann, 2011.

Practical component

Q1.Create a file “people.txt” with the following data:

Age	agegroup	height	status	yearsmarried
21	adult	6.0	single	-1
2	child	3	married	0
18	adult	5.7	married	20
221	elderly	5	widowed	2
34	child	-7	married	3

i) Read the data from the file "people.txt".

ii) Create a ruleset E that contains rules to check for the following conditions:

1. The age should be in the range 0-150.
2. The age should be greater than years married.
3. The status should be married, single, or widowed.

4. If age is less than 18, the age group should be child; if age is between 18 and 65, the age group should be adult; if age is more than 65, the age group should be elderly.

iii) Check whether ruleset E is violated by the data in the file people.txt.

iv) Summarize the results obtained in part (iii).

v) Visualize the results obtained in part (iii).

Q2. Perform the following preprocessing tasks on the dirty_iris dataset:

i) Calculate the number and percentage of observations that are complete.

ii) Replace all the special values in data with NA.

iii) Define these rules in a separate text file and read them:

- Species should be one of the following values: setosa, versicolor, or virginica.
- All measured numerical properties of an iris should be positive.
- The petal length of an iris is at least 2 times its petal width.
- The sepal length of an iris cannot exceed 30 cm.
- The sepals of an iris are longer than its petals.

Use the appropriate functions (e.g., editfile function in R with package editrules or similar function in Python) and print the resulting constraint object.

iv) Determine how often each rule is broken (violated edits). Also, summarize and plot the result.

v) Find outliers in sepal length using boxplot and boxplot.stats.

Q3. Load the data from the wine dataset. Check whether all attributes are standardized (mean is 0 and standard deviation is 1). If not, standardize the attributes. Do the same with the Iris dataset.

Q4. Run the Apriori algorithm to find frequent itemsets and association rules with the following parameters on any data set.

- Minimum support as 40% and minimum confidence as 80%.
- Minimum support as 50% and minimum confidence as 70%.

Q5. Use K-nearest neighbors algorithm to build classifiers. Divide the dataset into training and test sets and compare the accuracy of the different classifiers under the following situations:

a) Training set = 75%, Test set = 25%

b) Training set = 66.6% (2/3rd of total), Test set = 33.3%

The training set is chosen by: i) Holdout method

Q6. Use Decision tree classification algorithms to build classifiers. Divide the dataset into training and test sets and compare the accuracy of the different classifiers under the following situations:

a) Training set = 75%, Test set = 25%

b) Training set = 66.6% (2/3rd of total), Test set = 33.3%

The training set is chosen by: i) Random subsampling ii) Cross-Validation

Q7. Use the Simple K-means clustering algorithm to cluster the data. Compare the performance of clusters by changing the parameters involved in the algorithm.

Q8. Use the Hierarchical clustering algorithm to cluster the data. Compare the performance of clusters by changing the parameters involved in the algorithm.