Nirma University

Institute of Technology

Computer Engineering Department

Data Mining – 2CSDE71

B. Tech Semester – VI **Academic Year: Even 2023-2024**

List of Practical

Laboratory work will be based on the above syllabus with minimum 10 experiments to be incorporated.

<table>
<tr><th>Sr. No</th><th>List of Experiments</th><th>No. of Hours</th><th>Mapped CLO</th></tr>
<tr><td>1</td><td>Data Domain selection and Identification of Characteristics of selected Dataset of different Formats. Also write a report with following detail
a) Selection of data domain
b) Define the data domain
c) The data source
d) Objective
e) Define the selection of fields
f) Characteristic and behaviors (distribution and inference) of data for each selected field</td><td>2</td><td>CLO1</td></tr>
<tr><td>2</td><td>Calculate the dissimilarity value of the following table containing various types of attributes or on the dataset identified in Practical 1.
<table>
<tr><td></td><td>Nominal</td><td>Ordinal</td><td>Numerical</td></tr>
<tr><td>1</td><td>CodeA</td><td>Excellent</td><td>45</td></tr>
<tr><td>2</td><td>CodeB</td><td>Fair</td><td>22</td></tr>
<tr><td>3</td><td>CodeC</td><td>Good</td><td>64</td></tr>
<tr><td>4</td><td>CodeD</td><td>Excellent</td><td>28</td></tr>
</table></td><td>2</td><td>CLO1</td></tr>
<tr><td>3</td><td>Identify and implement any three methods to fill the missing values indicated by’?’ in the given data set or on the dataset identified in Practical 1. (Analysis of the best method with reasoning)
<table>
<tr><th>Name</th><th>Value</th></tr>
<tr><td>A</td><td>45</td></tr>
<tr><td>B</td><td>37</td></tr>
<tr><td>C</td><td>59</td></tr>
<tr><td>D</td><td>?</td></tr>
<tr><td>E</td><td>47</td></tr>
<tr><td>F</td><td>39</td></tr>
<tr><td>G</td><td>?</td></tr>
<tr><td>H</td><td>43</td></tr>
<tr><td>I</td><td>52</td></tr>
</table></td><td>2</td><td>CLO1</td></tr>
</table>

	J ?		
4	Implement the menu driven program for normalization for given dataset from user (by all means) for the following normalization techniques. for example, dataset (age): 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70. • Use Min-Max Normalization to transform the value 25 and 52 for age onto the range [0.0, 1.0]. • Use z-score normalization to transform the value 35 for *age*, where the standard deviation of *age* is 12.94 years. • Use normalization by decimal scaling to transform the value 35 for *age*.	2	CLO1
5	Implement Principal component analysis Dimensionality reduction technique on the following dataset (Do not use any available library of PCA)  	2	CLO2
6	Identify the frequent patterns and generate strong association rule from the frequent pattern for the following data set (using Apriori and any improved version of Apriori). Keep minimum 50% support and 40% confidence. | Tid | Items bought | |---|---| | 10 | Beer, Nuts, Diaper | | 20 | Beer, Coffee, Diaper | | 30 | Beer, Diaper, Eggs | | 40 | Nuts, Eggs, Milk | | 50 | Nuts, Coffee, Diaper, Eggs, Milk | * Prepare the analytics report on interesting patterns and correlation analysis of generated output.	2	CLO2

7	The DBLP dataset (www.informatik.uni-trier.de/ley/db/) consists of over one million entries of research papers published in computer science conferences and journals. Among these entries, there are a good number of authors that have coauthor relationships. (a) Propose a method and implement it to efficiently mine a set of coauthor relationships that are closely correlated (e.g., often coauthoring papers together).	4	CLO2
8	Implement C-Means clustering algorithm cluster the following four points (with (x; y) representing location) into two clusters. A1(1; 2), A2(2; 3), A3(9; 4), A4(10,1) B1(5; 8); B2(7; 5); B3(6; 4); C1(4; 2); C2(4; 9) The distance function is Euclidean distance. Suppose initially we assign A1, B1, and C1 as the centre of each cluster, respectively. Use the k-means algorithm to show the three cluster centres and all the points of clusters after the 2nd round of execution.	4	CLO2
9	Implement Random Forest Algorithm on Car Evaluation Database. It contains examples with the structural information removed, i.e., it directly relates CAR to the six input attributes: buying, maint, doors, persons, lug_boot, safety. Basically, we have to build a classifier to classify a car as 'Unacceptable', 'Acceptable', 'Good' and 'Very Good' based on the attributes. The different attributes values are given as follows: 1. **buying**: vhigh, high, med, low. 2. **maint**: vhigh, high, med, low. 3. **doors**: 2, 3, 4, 5, more. 4. **persons**: 2, 4, more. 5. **lug_boot**: small, med, big. 6. **safety**: low, med, high. You can download the dataset from here : https://archive.ics.uci.edu/ml/datasets/Car+Evaluation	4	CLO2 CLO3
10	Explore Weka/Knime tool for KDD process (classification algorithms) in any of the following application. 1. Healthcare 2. Government 3. Transportation 4. Education	4	CLO2 CLO3

	5. E-commerce 6. Entertainment		

Suggested Readings^:

1. Jiawei Han and Micheline Kamber, Data mining: Concepts and Techniques, Morgan Kaufmann Publishers.
2. Ian H. Witten and Eibe Frank, Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann
3. Hand, Mannila, and Smyth., Principles of Data Mining, MIT Press
4. Berry and Linoff, Mastering Data Mining, Wiley
5. Delmater and Hancock, Data Mining Explained, Digital Press

Lesson Plan

Sr. No.	Topics	CLO
1	Motivation and importance, different kinds of data, data mining functionalities,	1
2	classification of data mining systems,	1
3	major issues in data mining	1
4	Data summarization,	1
5	data cleaning,	1
6	data integration and transformation,	1
7	data reduction,	1
8	data discretization and concept hierarchy generation	1
9	Basic concept of Mining Frequent Patterns	2
10	efficient and scalable frequent itemset mining methods	2
11	efficient and scalable frequent itemset mining methods	2
12	Mining various kind of association rules	2
13	Mining various kind of association rules	2
14	from association mining to correlation analysis,	2
15	constraint-based association mining	2
16	Classification vs. prediction, Issues regarding classification and prediction	2
17	Classification by decision tree induction	2
18	Statistical-Based Algorithms, Distance-Based Algorithms	2
19	Decision Tree-Based Algorithms, Neural Network-Based Algorithms, Rule-Based Algorithms,	2
20	Combining Techniques, accuracy and error measures, evaluation of the accuracy of a classifier or predictor.	2
21	Types of data in cluster analysis	2,3
22	overview of major clustering methods,	2,3
23	Probabilistic model based clustering	2,3

24	Clustering high dimensional data,	2,3
25	Clustering Graph and Network data.	2,3
26	Applications of Distributed and parallel Data Mining	2,3
27	Advanced Techniques: Web Mining	2,3
28	Spatial Database Mining	2,3
29	Temporal Mining,	2,3
30	Multimedia Mining.	2,3