| | | | |
|---|---|---|---|
| | **Hope Foundation's**<br>**Finolex Academy of Management and Technology, Ratnagiri** | | |
| | **Department of Computer Science and Engineering (AIML)** | | |
| Subject name: Data warehousing and Mining Lab | | | Subject Code: CSL503 |
| Class | TE CSE | Semester –V (CBCGS) | Academic year: 2024-25 |
| Name of Student | | | **QUIZ Score :** |
| Roll No | | Experiment No. | 04 |
| Title: **Using open source tools perform Clustering.** | | | |

**1. Lab objectives applicable:**

**LOB4:** To make students well versed in all data mining algorithms, methods, and tools.

**2. Lab outcomes applicable:**

**LO3:** Demonstrate an understanding of the importance of data mining.

**LO6:** Implement the appropriate data mining methods like classification, clustering or Frequent Pattern mining on large data sets.

**3. Learning Objectives:**

1. To determine similarity and dissimilarity among elements and create clusters accordingly.

**4. Practical applications of the assignment/experiment:**

Clustering algorithms group similar data points together to uncover patterns and relationships, enhancing data analysis and decision-making.

**5. Prerequisites**:

NA

**6. Minimum Hardware Requirements**:

1. I series processor, RAM 4GB,

**7. Software Requirements:**

1. Weka 3.8

**8. Quiz Questions**

https://docs.google.com/forms/d/e/1FAIpQLSds1ANl3PsFjWngRtdL9p9QtcRKBfZ1sGM4s6yQD3hGHG9olQ/viewform?usp=sf_link

**9. Experiment/Assignment Evaluation:**

| **Sr. No.** | **Parameters** | **Marks obtained** | **Out of** |
|---|---|---|---|
| **1** | Technical Understanding (Assessment may be done based on Q & A **or** any other relevant method.) Teacher should mention the other method used - | | 6 |
| **2** | Lab Performance | | 2 |
| **3** | Punctuality | | 2 |
| **Date of performance (DOP)** | | **Total marks obtained** | **10** |

**Signature of Faculty**

**11. Installation Steps / Performance Steps and Results –**

**Source code:**

1.kmeans_1D:

Clusterer output

```
Attributes:   1
              quantity
Test mode:    evaluate on training data


=== Clustering model (full training set) ===


kMeans
======

Number of iterations: 4
Within cluster sum of squared errors: 0.1913265306122449

Initial starting points (random):

Cluster 0: 2
Cluster 1: 12

Missing values globally replaced with mean/mode

Final cluster centroids:
                        Cluster#
Attribute    Full Data          0          1
                (9.0)      (6.0)      (3.0)
================================================
quantity            13          7         25



Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       6 ( 67%)
1       3 ( 33%)
```

Viewer

Relation: kmeans_1D

| No. | 1: quantity Numeric |
|---|---|
| 1 | 2.0 |
| 2 | 4.0 |
| 3 | 10.0 |
| 4 | 12.0 |
| 5 | 3.0 |
| 6 | 20.0 |
| 7 | 30.0 |
| 8 | 11.0 |
| 9 | 25.0 |

Add instance | Undo | OK

2.k_means2D:

Clusterer output

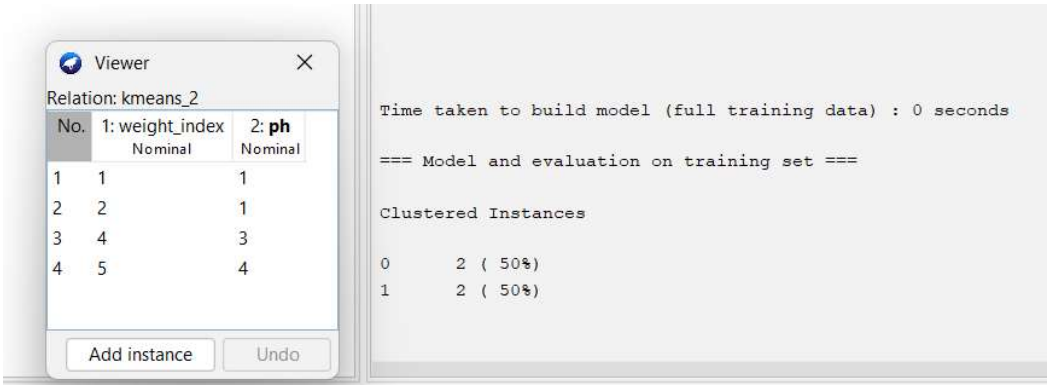
```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2
Relation:     kmeans_2
Instances:    4
Attributes:   2
              weight_index
              ph
Test mode:    evaluate on training data


=== Clustering model (full training set) ===


kMeans
======

Number of iterations: 2
Within cluster sum of squared errors: 3.0000000000000004

Initial starting points (random):

Cluster 0: 4,3
Cluster 1: 1,1

Missing values globally replaced with mean/mode

Final cluster centroids:
                        Cluster#
Attribute       Full Data          0          1
                    (4.0)      (2.0)      (2.0)
================================================
weight_index            1          4          1
ph                      1          3          1
```

Viewer

Relation: kmeans_2

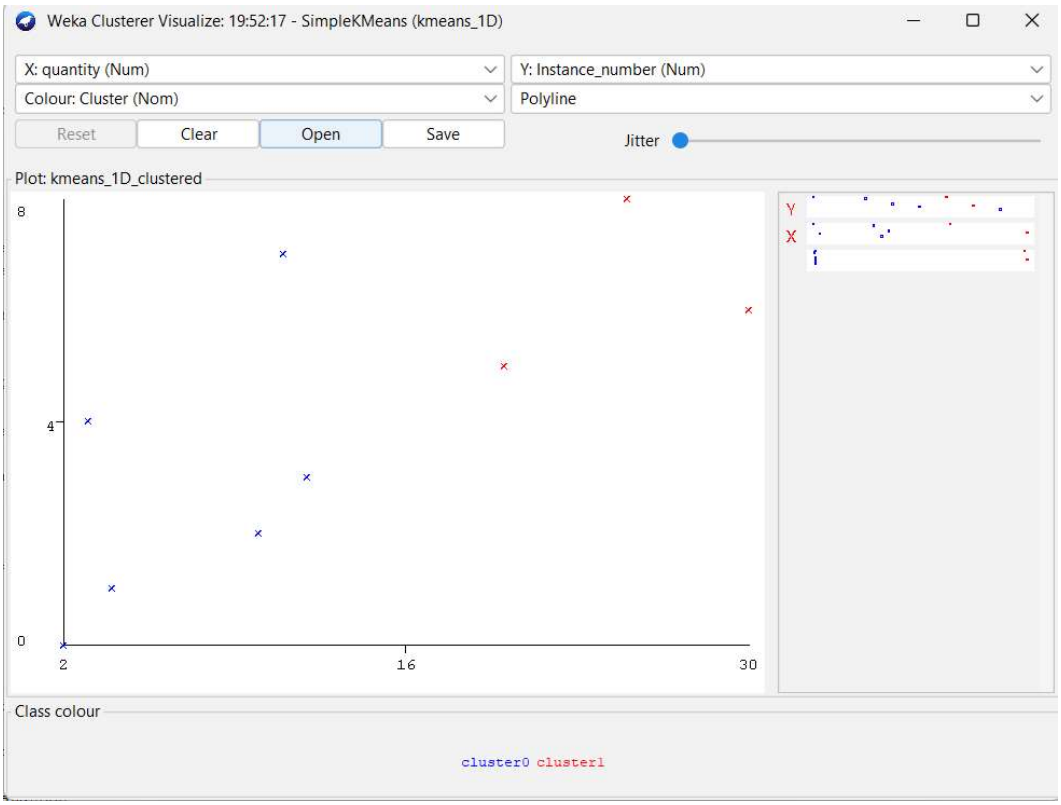
| No. | 1: weight_index Nominal | 2: ph Nominal |
|---|---|---|
| 1 | 1 | 1 |
| 2 | 2 | 1 |
| 3 | 4 | 3 |
| 4 | 5 | 4 |

Add instance | Undo

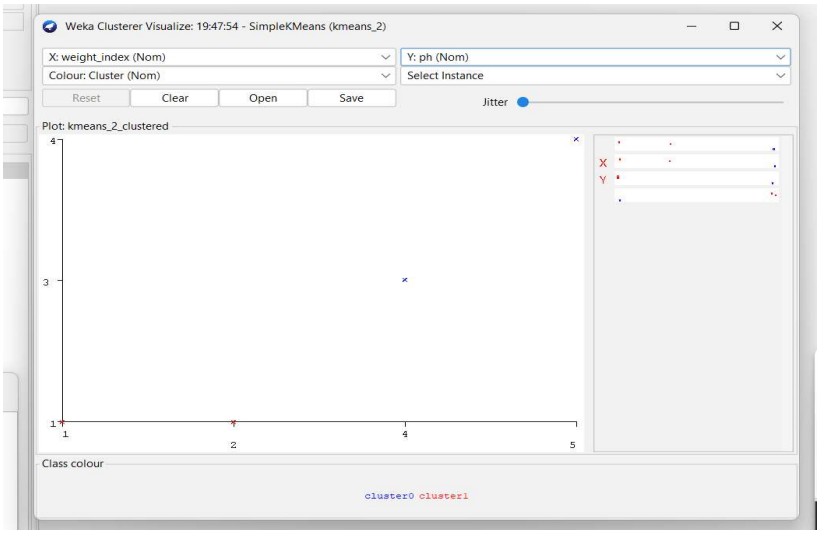



**Output:**

1.kmeans_1D:




**2.**kmeans_2D:



**12. Learning Outcomes Achieved**

1. Students are able to cluster the given data in k- some known number of clusters.

**13. Conclusion:**

**1. Applications of the Studied Technique in Industry**

Clustering algorithms, such as K-means or hierarchical clustering, are widely used in industry for customer segmentation, market analysis, and anomaly detection. These techniques help businesses tailor marketing strategies, optimize resource allocation, and identify unusual patterns or trends in large datasets

**2. Engineering Relevance**

Clustering algorithms are crucial in engineering for solving complex problems related to pattern recognition, image processing, and system optimization. They enable engineers to group similar data points, improve model accuracy, and make informed decisions based on data-driven insights.

**3. Skills Developed**

The experiment with clustering algorithms enhances skills in data preprocessing, algorithm implementation, and result interpretation. It also develops expertise in applying statistical techniques to solve real-world problems, as well as proficiency in using data mining tools and software for effective data analysis.

**14. References**:

[1] https:// Paulraj Ponniah, "Data Warehousing: Fundamentals for IT Professional" , Wiley Publications
[2] Han, Kamber, "Data Mining Concepts and Techniques", Morgan Kaufmann 3nd Edition.
[3] Margaret H. Dunham, "Data Mining: Introductory and Advanced Topics", Person Education.
[4] Raghu Ramakrishnan and Johannes Gehrke, "Database Management Systems", 3rd Edition McGraw Hill.
[5] Elmasari and Navathe, "Fundamentals of Database Systems", Pearson Education.