Experiment - 3.2

Student Name: Pankaj Singh Kanyal **UID:** 20BCS6668
Branch: AIML **Section/Group:** 20AML-4B
Semester: 5 **Date:**
Subject Name: Data mining & warehousing lab **Subject Code:** 20CSF-333_20AML-4

1. **Aim/Overview of the Practical**

Write a procedure for Employee data using Make Density Based Cluster Algorithm.

2. **Task to be done**

1) Create the weather table using notepad
2) Create data and use the evaluation and visualize tab to flow the data using knowledge flow.

3. **Program Code:**

Theory
Cluster analysis or clustering is the task of assigning a set of objects into groups (called clusters) so that the objects in the same cluster are more similar (in some sense or another) to each other than to those in other clusters. Clustering is the main task of explorative data mining, and a common technique for statistical data analysis used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics.

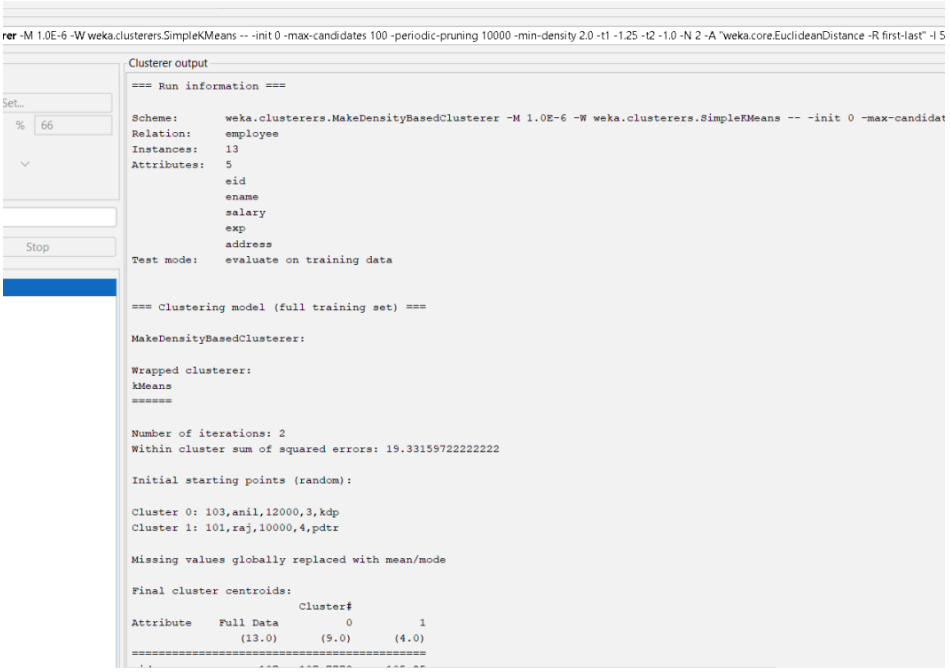
3. Steps for procedure:
Creation of Customer Table:
1. Open Start -> Programs -> Accessories -> Notepad.
2. Type the following training data with the help of Notepad for Employee table.
3. After that save the file with .arff file format.

Procedure:
1) Click Start -> Programs -> Weka 3.4
2) Click on Explorer.
3) Click on open file & then select Employee.arff file.
4) Click on Cluster menu. In this there are different algorithms are there.
5) Click on Choose button and then select MakeDensityBasedClusterer algorithm.
6) Click on Start button and then output will be displayed on the screen.

Data used in arfff file

@relation employee
@attribute eid numeric
@attribute ename {raj,ramu,anil,sunil,rajiv,sunitha,kavitha,suresh,ravi,ramana,ram,kavya,navya}
@attribute salary numeric
@attribute exp numeric
@attribute address {pdtr,kdp,nlr,gtr}
@data 101,raj,10000,4,pdtr
102,ramu,15000,5,pdtr
103,anil,12000,3,kdp
104,sunil,13000,3,kdp
105,rajiv,16000,6,kdp
106,sunitha,15000,5,nlr
107,kavitha,12000,3,nlr
108,suresh,11000,5,gtr
109,ravi,12000,3,gtr
110,ramana,11000,5,gtr
111,ram,12000,3,kdp
112,kavya,13000,4,kdp
113,navya,14000,5,kdp

4. **Output**



```
rer -M 1.0E-6 -W weka.clusterers.SimpleKMeans -- -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 5
```

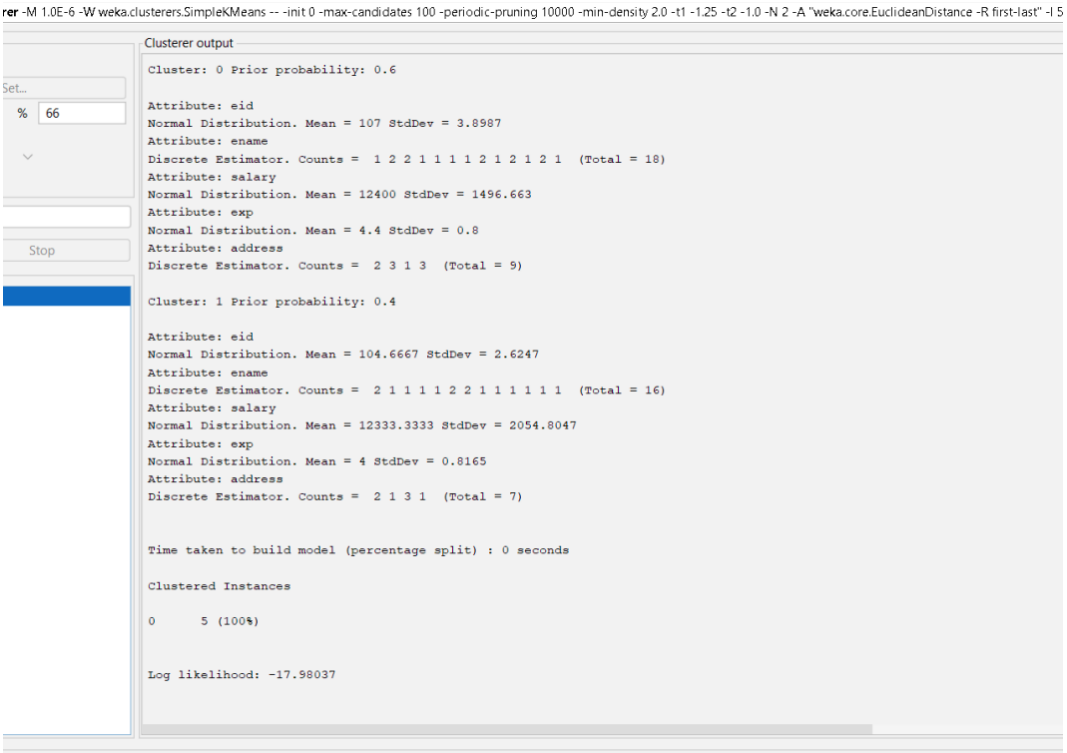
Clusterer output

```
Normal Distribution. Mean = 107.7778 StdDev = 3.4247
Attribute: ename
Discrete Estimator. Counts =  1 1 2 2 2 2 2 1 2 1 2 2 2  (Total = 22)
Attribute: salary
Normal Distribution. Mean = 13222.2222 StdDev = 1396.645
Attribute: exp
Normal Distribution. Mean = 3.8889 StdDev = 1.0999
Attribute: address
Discrete Estimator. Counts =  1 7 3 2  (Total = 13)

Cluster: 1 Prior probability: 0.3333

Attribute: eid
Normal Distribution. Mean = 105.25 StdDev = 3.8324
Attribute: ename
Discrete Estimator. Counts =  2 2 1 1 1 1 1 2 1 2 1 1 1  (Total = 17)
Attribute: salary
Normal Distribution. Mean = 11750 StdDev = 1920.2864
Attribute: exp
Normal Distribution. Mean = 4.75 StdDev = 0.433
Attribute: address
Discrete Estimator. Counts =  3 1 1 3  (Total = 8)


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

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

Clustered Instances

0       9 ( 69%)
1       4 ( 31%)


Log likelihood: -16.52967
```

rer -M 1.0E-6 -W weka.clusterers.SimpleKMeans -- -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 5

Clusterer output

```
Final cluster centroids:
                                Cluster#
Attribute    Full Data          0          1
                (13.0)       (9.0)      (4.0)
=============================================
eid                107    107.7778     105.25
ename              raj         anil        raj
salary       12769.2308 13222.2222      11750
exp             4.1538      3.8889       4.75
address            kdp         kdp       pdtr


Fitted estimators (with ML estimates of variance):

Cluster: 0 Prior probability: 0.6667

Attribute: eid
Normal Distribution. Mean = 107.7778 StdDev = 3.4247
Attribute: ename
Discrete Estimator. Counts =  1 1 2 2 2 2 2 1 2 1 2 2 2  (Total = 22)
Attribute: salary
Normal Distribution. Mean = 13222.2222 StdDev = 1396.645
Attribute: exp
Normal Distribution. Mean = 3.8889 StdDev = 1.0999
Attribute: address
Discrete Estimator. Counts =  1 7 3 2  (Total = 13)

Cluster: 1 Prior probability: 0.3333

Attribute: eid
Normal Distribution. Mean = 105.25 StdDev = 3.8324
Attribute: ename
Discrete Estimator. Counts =  2 2 1 1 1 1 1 2 1 2 1 1 1  (Total = 17)
Attribute: salary
Normal Distribution. Mean = 11750 StdDev = 1920.2864
Attribute: exp
Normal Distribution. Mean = 4.75 StdDev = 0.433
```



5. Result and Conclusion

Successfully implemented the Density Based Cluster Algorithm for employee data.

6. Learning Outcomes

1. Learned to use knowledge flow in WEKA data mining software
2. Learned about Arff loaders and use them in knowledge flow
3. Learned and implemented cross validation techniques in weather dataset.

Evaluation Grid (To be created as per the SOP and Assessment guidelines by the faculty):

Sr. No.	Parameters	Marks Obtained	Maximum Marks
1.			
2.			
3.			