Experiment No: 08	TE AI&DS
Date of Performance:	Roll No: 9696
Aim: : To analyze and evaluate the performance of different Clustering techniques using WEKA tool	
Related CO5: To analyze and evaluate perform of data mining techniques applied on large dataset using open-source tool for data mining	
Objective: To learn WEKA(Data Mining tool) and analyze and evaluate clustering algorithm performance.	

Rubrics for assessment of Experiment:

Sr. No	**Parameters**	**Exceed Expectations(EE)**	**Meet Expectations (ME)**	**Below Expectations (BE)**
1	Timeline (2)	Early or on time (2)	One session late (1)	More than one session late (0)
2	Preparedness (2)	Knows the basic theory related to the experiment very well. (2)	Managed to explain the theory related to the experiment. (1)	Not aware of the theory to the point. (1)
3	Effort (3)	Done expt on their own. (3)	Done expt with help from other. (2)	Just managed. (1)
4	Documentation(2)	Lab experiment is documented in proper format and maintained neatly. (2)	Documented in proper format but some formatting guidelines are missed. (1)	Experiments not written in proper format (0.5)
5	Result (1)	Specific conclusion.(1)	Partially specific conclusion. (0.5)	Not specific at all. (0)

Assessment Marks:

Timeline(2)	Preparedness(2)	Effort(3)	Documentation(2)	Result(1)	Total(10)

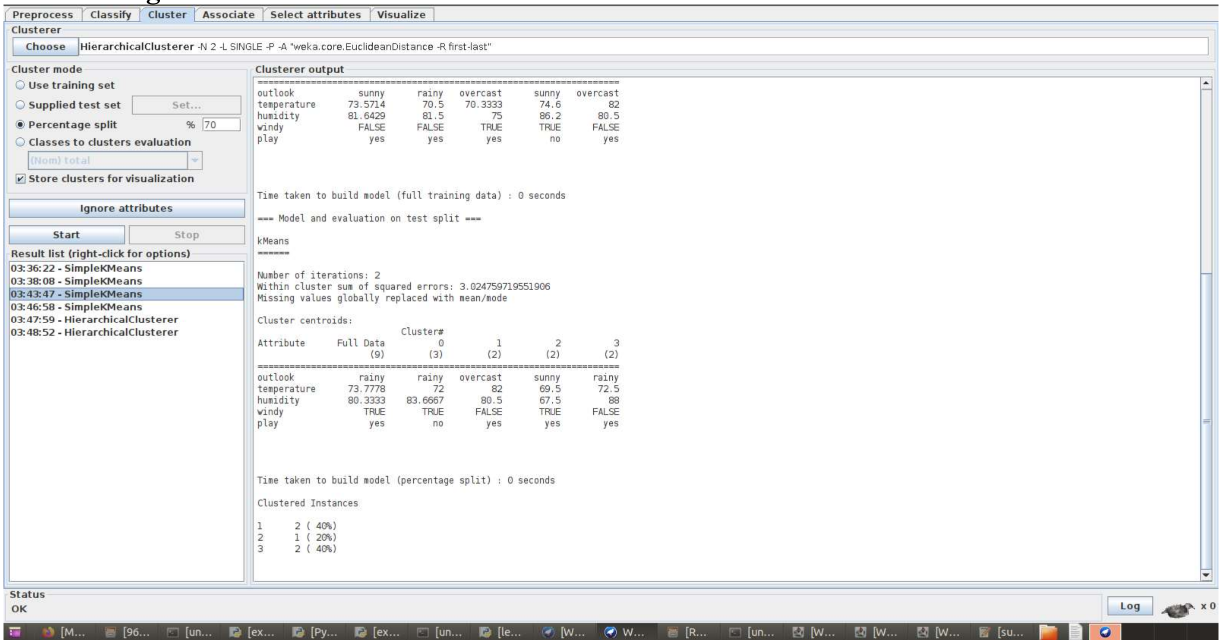
Theory:

WEKA is a data mining system developed by the University of Waikato in New Zealand that implements data mining algorithms. It is a collection of machine learning algorithms for data mining tasks. The algorithms are applied directly to a dataset.
WEKA implements algorithms for data preprocessing, classification, regression, clustering, association rules; it also includes a visualization tools. When 'WEKA GUI Chooser' window appears on the screen, we can select one of the four options.
The options are 1. Simple CLI : provides a simple command line interface and allows direct execution of WEKA commands. 2. Explorer : is an environment for exploring data. 3. Experimenter : is an environment for performing experiments and conducting statistical tests between learning schemes. 4. Knowledge Flow : is a Java-Beans-based interface for setting up and running machine learning experiments. Classifiers in WEKA are the models for predicting nominal or numeric quantities. The learning schemes available in WEKA include decision trees, Bayes net, neural network, support vector machine and so on. In this experiment analysis and evaluation of three classification algorithm: Naive Bayesian algorithm, C4.5 algorithm, zero R is done. Before running the classification algorithm it is required to set test options. The selected test options were : 1. Use training set : Evaluates the classifier on how well it predicts the class of the instances it was trained on. 2. Cross-validation : Evaluates the classifier by cross-validation, using the number of folds (10) 3. Percentage split : Evaluates the classifier on how well it predicts a certain percentage of the data, which is held out for testing. In the classifier evaluation options following options are checked : 1. Output model : The output is the classification model on the full training set. 2. Output per-class stats : The precision/recall and true/false statistics for each class output 3. Output confusion matrix : The confusion matrix of one classifier's prediction is included in the output When training set is complete, the 'classifier' output area on the right panel of the classifier window is filled with text describing the results of training and testing.

Practical Exercise:

Apply and evaluate the result for different classification techniques on various datasets using WEKA.

K Mean Algorithm



HierarchicalClusterer -N 2 -L SINGLE -P -A "weka.core.EuclideanDistance -R first-last"

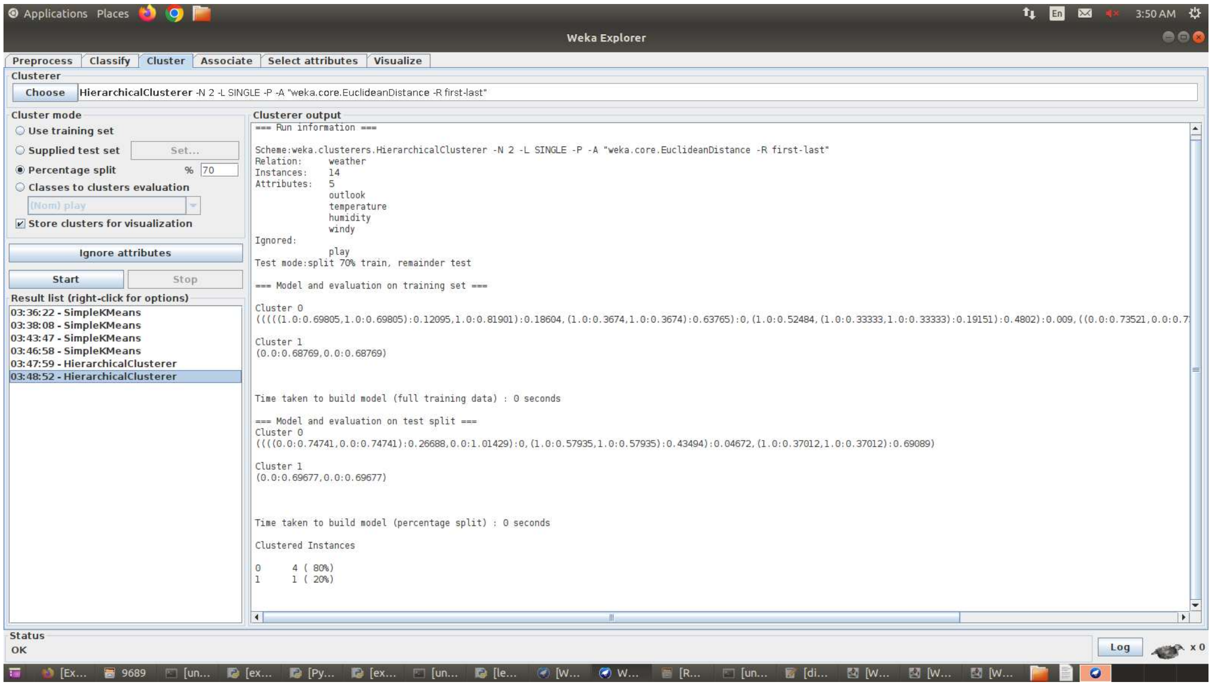
Percentage split % 70

```
03:36:22 - SimpleKMeans
03:38:08 - SimpleKMeans
03:43:47 - SimpleKMeans
03:46:58 - SimpleKMeans
03:47:59 - HierarchicalClusterer
03:48:52 - HierarchicalClusterer
```

```
=====================================================================
outlook             sunny      rainy   overcast      sunny   overcast
temperature       73.5714       70.5    70.3333       74.6         82
humidity          81.6429       81.5         75       86.2       80.5
windy               FALSE      FALSE       TRUE       TRUE      FALSE
play                  yes        yes        yes         no        yes




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

=== Model and evaluation on test split ===

kMeans
======

Number of iterations: 2
Within cluster sum of squared errors: 3.024759719551906
Missing values globally replaced with mean/mode

Cluster centroids:
                         Cluster#
Attribute     Full Data         0          1          2          3
                    (9)       (3)        (2)        (2)        (2)
=====================================================================
outlook             rainy      rainy   overcast      sunny      rainy
temperature       73.7778         72         82       69.5       72.5
humidity          80.3333    83.6667       80.5       67.5         88
windy                TRUE       TRUE      FALSE       TRUE      FALSE
play                  yes         no        yes        yes        yes




Time taken to build model (percentage split) : 0 seconds

Clustered Instances

1      2 ( 40%)
2      1 ( 20%)
3      2 ( 40%)
```

Status
OK

Hierarchical Clustring:



Clusterer: HierarchicalClusterer -N 2 -L SINGLE -P -A "weka.core.EuclideanDistance -R first-last"

Cluster mode: ● Percentage split % 70

☑ Store clusters for visualization

Result list (right-click for options)

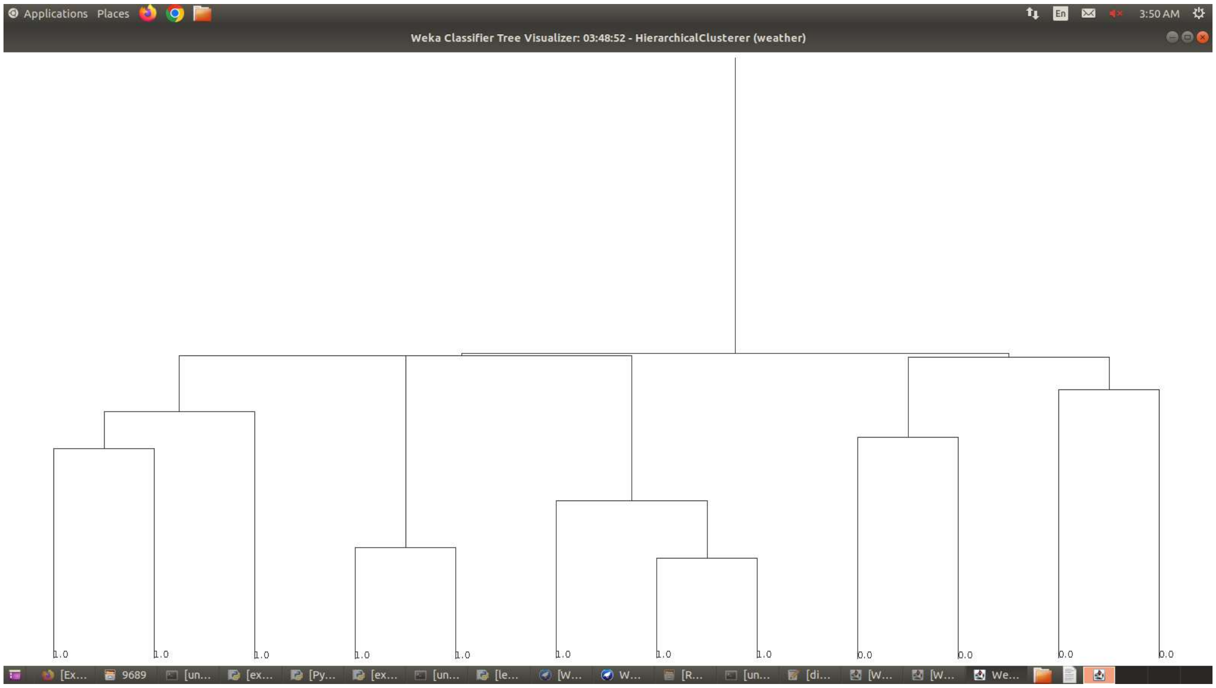
```
03:36:22 - SimpleKMeans
03:38:08 - SimpleKMeans
03:43:47 - SimpleKMeans
03:46:58 - SimpleKMeans
03:47:59 - HierarchicalClusterer
03:48:52 - HierarchicalClusterer
```

Clusterer output

```
=== Run information ===

Scheme:weka.clusterers.HierarchicalClusterer -N 2 -L SINGLE -P -A "weka.core.EuclideanDistance -R first-last"
Relation:     weather
Instances:    14
Attributes:   5
              outlook
              temperature
              humidity
              windy
Ignored:
              play
Test mode:split 70% train, remainder test

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

Cluster 0
(((((1.0:0.69805,1.0:0.69805):0.12095,1.0:0.81901):0.18604,(1.0:0.3674,1.0:0.3674):0.63765):0,(1.0:0.52484,(1.0:0.33333,1.0:0.33333):0.19151):0.4802):0.009,((0.0:0.73521,0.0:0.7

Cluster 1
(0.0:0.68769,0.0:0.68769)



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

=== Model and evaluation on test split ===
Cluster 0
((((0.0:0.74741,0.0:0.74741):0.26688,0.0:1.01429):0,(1.0:0.57935,1.0:0.57935):0.43494):0.04672,(1.0:0.37012,1.0:0.37012):0.69089)

Cluster 1
(0.0:0.69677,0.0:0.69677)



Time taken to build model (percentage split) : 0 seconds

Clustered Instances

0      4 ( 80%)
1      1 ( 20%)
```

Status
OK

Dendogram:

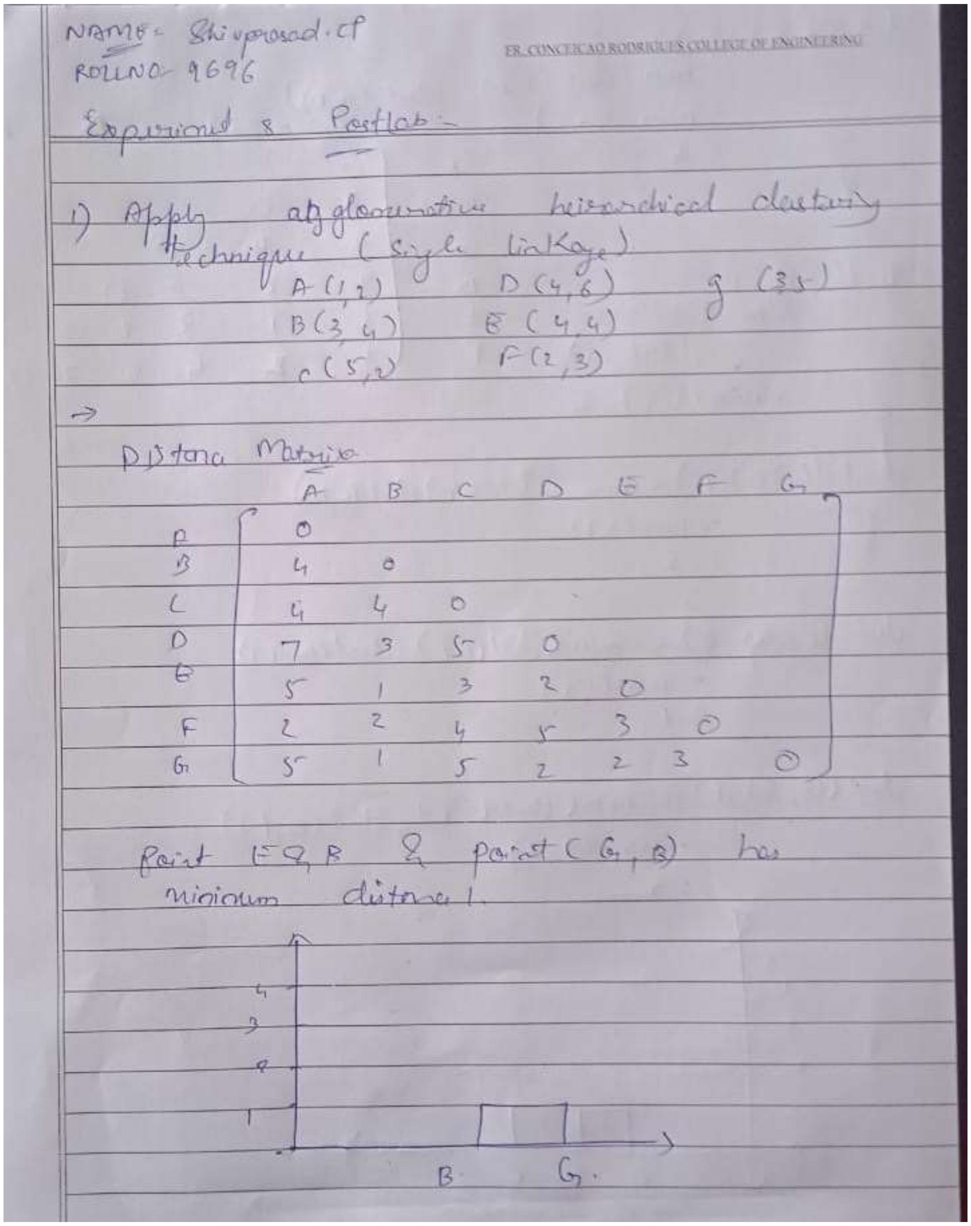


Applications
Places
3:50 AM
Weka Classifier Tree Visualizer: 03:48:52 - HierarchicalClusterer (weather)
1.0
1.0
1.0
1.0
1.0
1.0
1.0
1.0
0.0
0.0
0.0
0.0

Post lab Questions:

1. Apply agglomerative hierarchical clustering technique (single linkage) on following dataset.

   A(1,2)
   B(3,4)
   C(5,2)
   D(4,6)
   E(4,4)
   F(2,3)
   G(3,5)



NAME- Shivprasad.cP
ROLLNO- 9696

FR. CONCEICAO RODRIGUES COLLEGE OF ENGINEERING

Experiment 8 Postlab:-

1) Apply agglomerative heirarchical clustering technique (single linkage)

A (1,2) D (4,6) g (3,5)
B (3,4) E (4,4)
c (5,2) F (2,3)

→

Distance Matrix

$$
\begin{array}{c|ccccccc}
 & A & B & C & D & E & F & G \\
\hline
A & 0 & & & & & & \\
B & 4 & 0 & & & & & \\
C & 4 & 4 & 0 & & & & \\
D & 7 & 3 & 5 & 0 & & & \\
E & 5 & 1 & 3 & 2 & 0 & & \\
F & 2 & 2 & 4 & 5 & 3 & 0 & \\
G & 5 & 1 & 5 & 2 & 2 & 3 & 0
\end{array}
$$

Point E & B & point (G, B) has minimum distance 1.

$dist((B,G), A) =$
$\quad \min(dist(B,A), dist(G,A))$
$\quad = \min(4,5)$
$\quad = 4.$

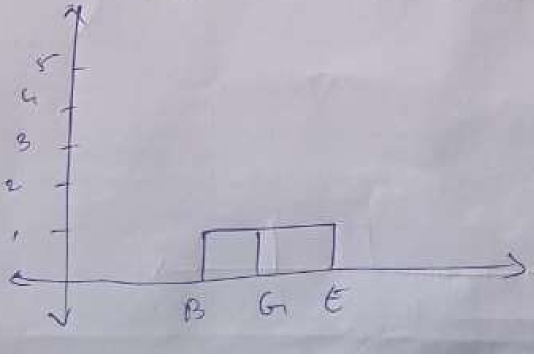
$$\begin{array}{c|cccccc} & A & (B,G) & c & D & e & F \\ \hline A & 0 & & & & & \\ (B,G) & 4 & 0 & & & & \\ G & 4 & 4 & 0 & & & \\ D & 7 & 2 & 5 & 0 & & \\ E & 5 & 1 & 3 & 2 & 0 & \\ F & 2 & 2 & 4 & 5 & 3 & 0 \end{array}$$

$dist((B,G), C)$
$\quad = \min(dist(B,C),$
$\quad\quad dist(G,C))$
$\quad = \min(4,5) = 4.$

$dist((B,G), D) = \min(dist(B,D)\ dist(G,D))$
$\quad = \min(3,2)$
$\quad = 2.$

$dist((B,G), e) = \min(dist(B,E), dist(G,E))$
$\quad = \min(1,2)$
$\quad = 1.$

$dist((B,G), F) = \min(dist(B,F), dist(G,F))$
$\quad = \min(2,3)$
$\quad = 2.$



$$\begin{array}{c|ccccc} & A & (B,G),E) & C & D & F \\ \hline A & 0 & & & & \\ ((B,G),E) & 4 & 0 & & & \\ C & 4 & 3 & 0 & & \\ D & 7 & 2 & 5 & 0 & \\ F & 2 & 2 & 4 & 5 & 0 \end{array}$$

$\text{dist}(((B,G),E),A).$

$= \min(\text{dist}(B,A), \text{dist}(G,A), \text{dist}(E,A)))$

$= \min(4,5,5)$

$= 4.$

$\text{dist}(((B,G),E),C) = \min(\text{dist}(B,C), \text{dist}(G,C), \text{dist}(E,C))$

$= \min(4,5,3)$

$= 3.$

$\text{dist}(((B,G),E),D) = \min(\text{dist}(B,D), \text{dist}(G,D), \text{dist}(E,D))$

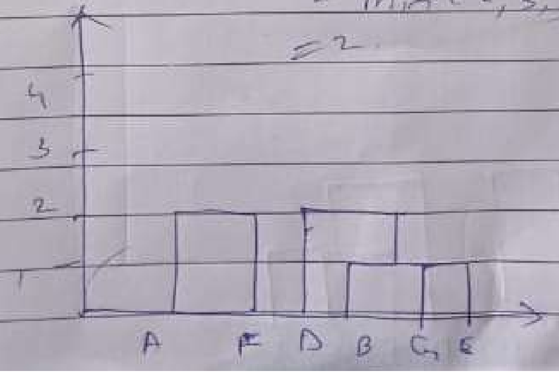
$= \min(3,4,2)$

$= 2$

$\text{dist}(((B,G),E),F) = \min(\text{dist}(B,F), \text{dist}(G,F), \text{dist}(E,F))$

$= \min(2,3,3)$

$= 2.$



dist ((A,F), C)

= min (dist (A,C) , dist (F,C))

= min (4, 2)

= 2.

$$\begin{array}{c|ccc} & (A,F) & C & ((B,G),E),D) \\ \hline (A,F) & 0 & & \\ C & 2 & 0 & \\ ((B,G,E),D) & 2 & 3 & 0 \end{array}$$

dist (A,F,(((B,G),E),D))

= dist min (dist (A,D) , dist (F,D), dist (A,B), dist (A,G) , dist (A,E), dist (F,B) dist (F,G) , dist (F,E))

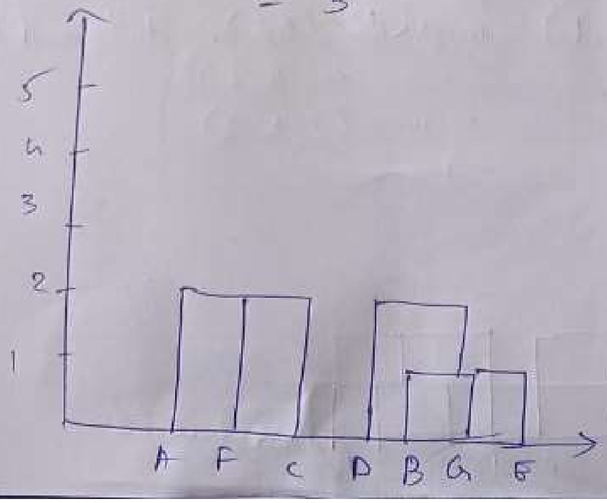
= min (7, 5, 4, 5, 5, 2, 3, 3)

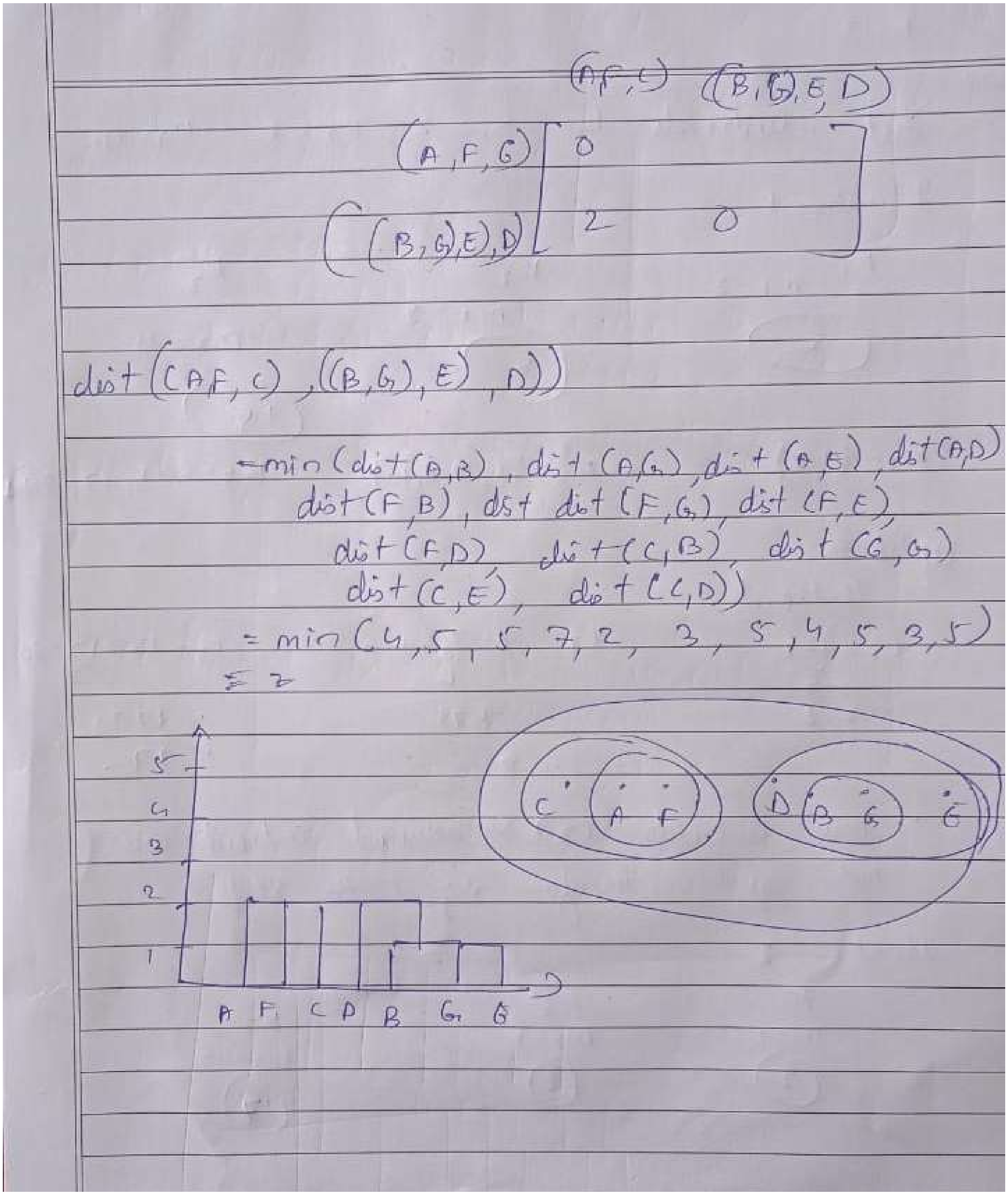
= 2.

dist (C, (((B,G),E),D)) = min (dist (C,D), dist (CB), dist (C,G) dot (C,E)).

= min (5, 4, 5, 3)

= 3.





$$\begin{array}{c|cc} & (A,F,C) & ((B,G),E,D) \\ \hline (A,F,G) & 0 & \\ (((B,G),E),D) & 2 & 0 \end{array}$$

dist((A,F,C),((B,G),E),D))

= min(dist(A,B), dist(A,G), dist(A,E), dist(A,D)
dist(F,B), dst dist(F,G), dist(F,E),
dist(F,D), dist(C,B), dist(G,G)
dist(C,E), dist(C,D))

= min(4,5,5,7,2,3,5,4,5,3,5)

= 2

5
4
3
2
1
A
F
C
D
B
G
E
C
A
F
D
B
G
E

2. Apply k-mean clustering technique on following dataset.
   1,2,6,7,8,10,15,17,20
   K=3

2) 1, 2, 6, 7, 8, 10, 15, 17, 20

K = 3

→

Let centroids be $K_1 = 2$, $K_2 = 8$, $K_3 = 15$

Iteration 1

K_1	K_2	K_3
$m_1 = 2$	$m_2 = 8$	$m_3 = 15$
2	8	15

$K_1 = \{1, 2\}$ $K_2 = \{6, 7, 8, 10\}$, $K_3 = \{15, 17, 20\}$

update the centroid.

$C_1 = \frac{1+2}{2}$ $= 2.$

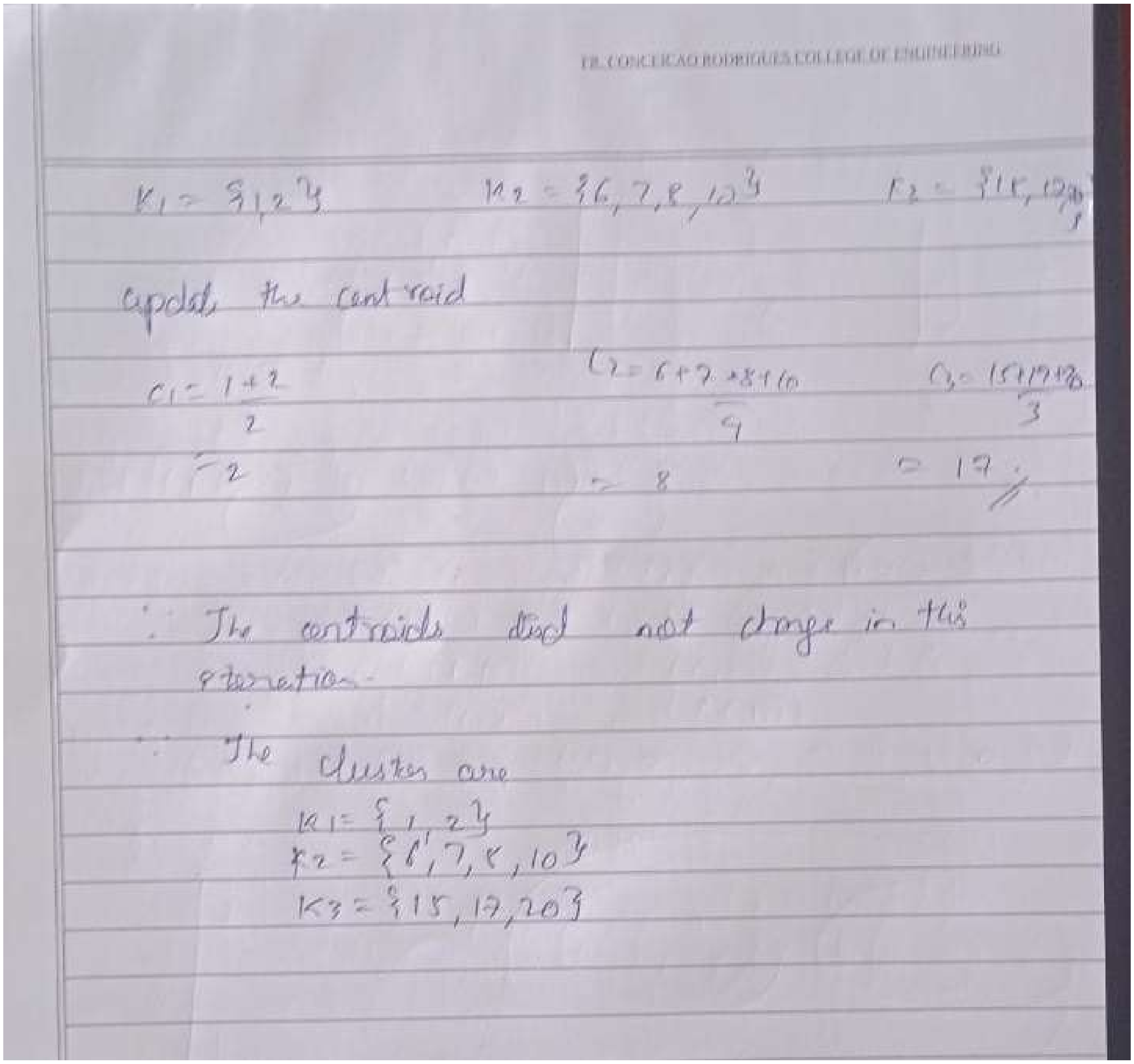
$C_2 = \frac{6+7,8,+10}{4}$ $= 7.75$ ≈ 8

$C_3 = \frac{15+17+20}{3}$ $= 17.33$ ≈ 17

∵ new centroids C(3) & assumed centroid are not equal, therefore repeat the step.

Iteration 2

K_1	K_2	K_3
$m_1 = 2$	$m_2 = 8$	$m_3 = 17$
2	8	17



FR. CONCEICAO RODRIGUES COLLEGE OF ENGINEERING

$K_1 = \{1,2\}$ $K_2 = \{6,7,8,10\}$ $K_3 = \{15, 17, 20\}$

updat the cent roid

$c_1 = \frac{1+2}{2}$ $C_2 = \frac{6+7+8+10}{4}$ $C_3 = \frac{15+17+20}{3}$

$= 2$ $= 8$ $= 17$ //

∴ The centroids did not change in this iteration.

∴ The clusters are

$K_1 = \{1, 2\}$

$K_2 = \{6, 7, 8, 10\}$

$K_3 = \{15, 17, 20\}$

Conclusion:

We are able to understand different clustering algorithm and am able to implement the algorithm using WEKA tool.