



Data Mining

CS475

Spring 2019

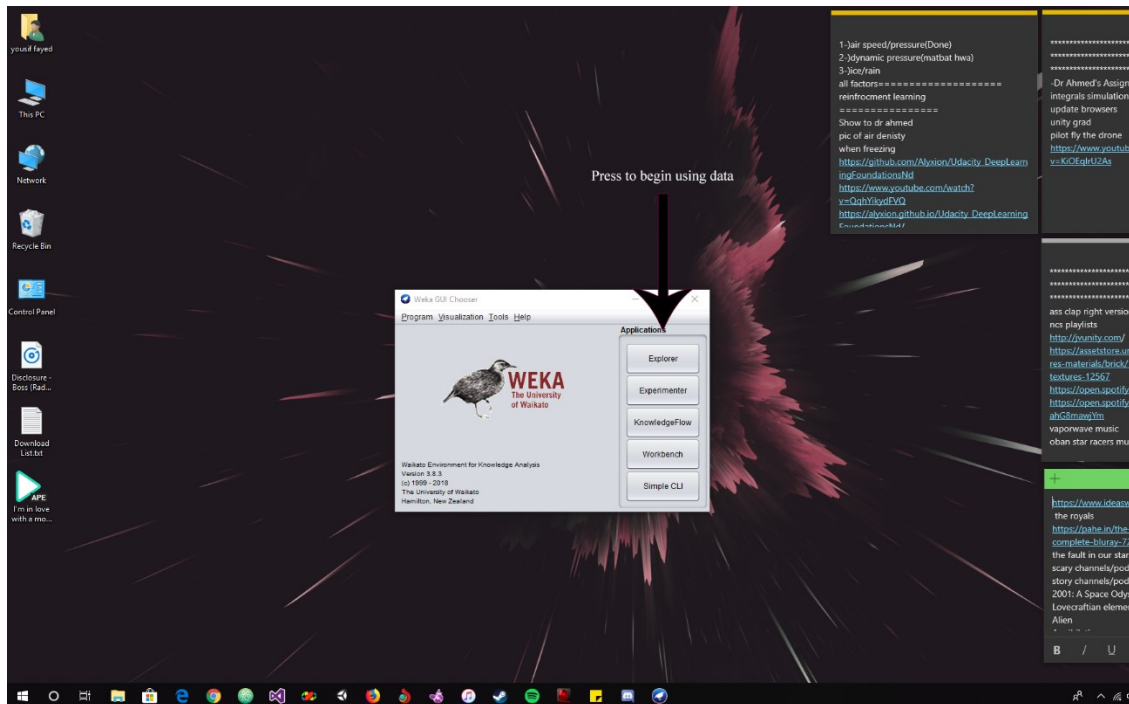
Project

Name: Youssef Ali Fayed

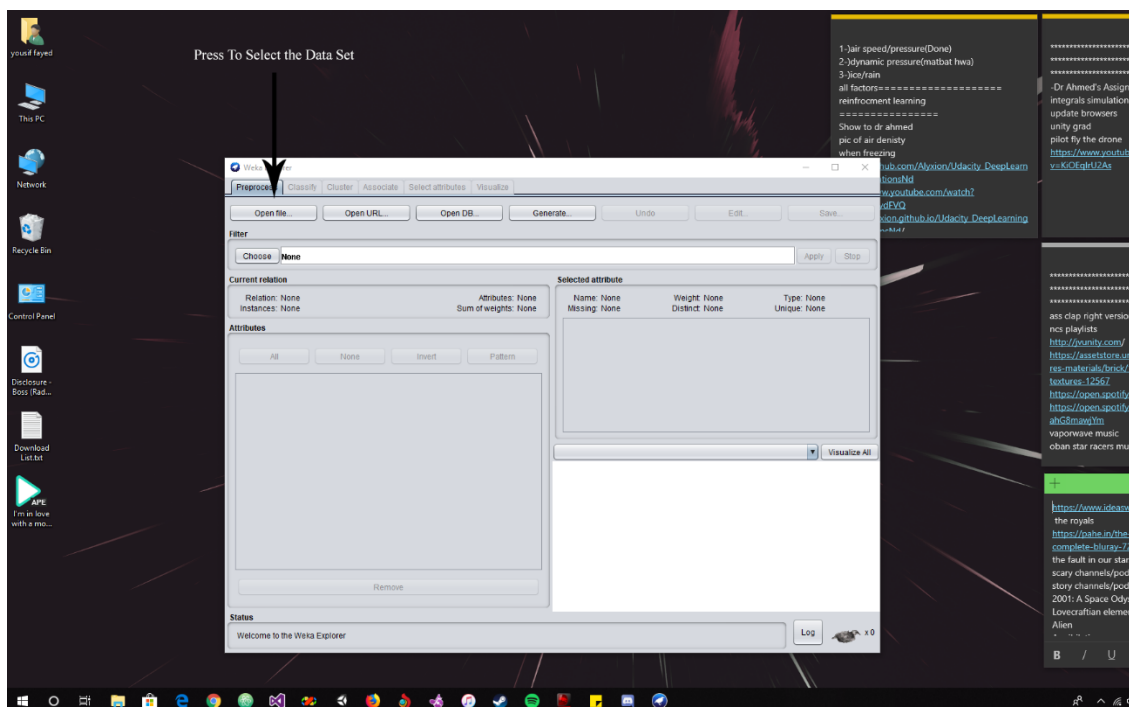
ID: 162085

1-)Open Weka Program.

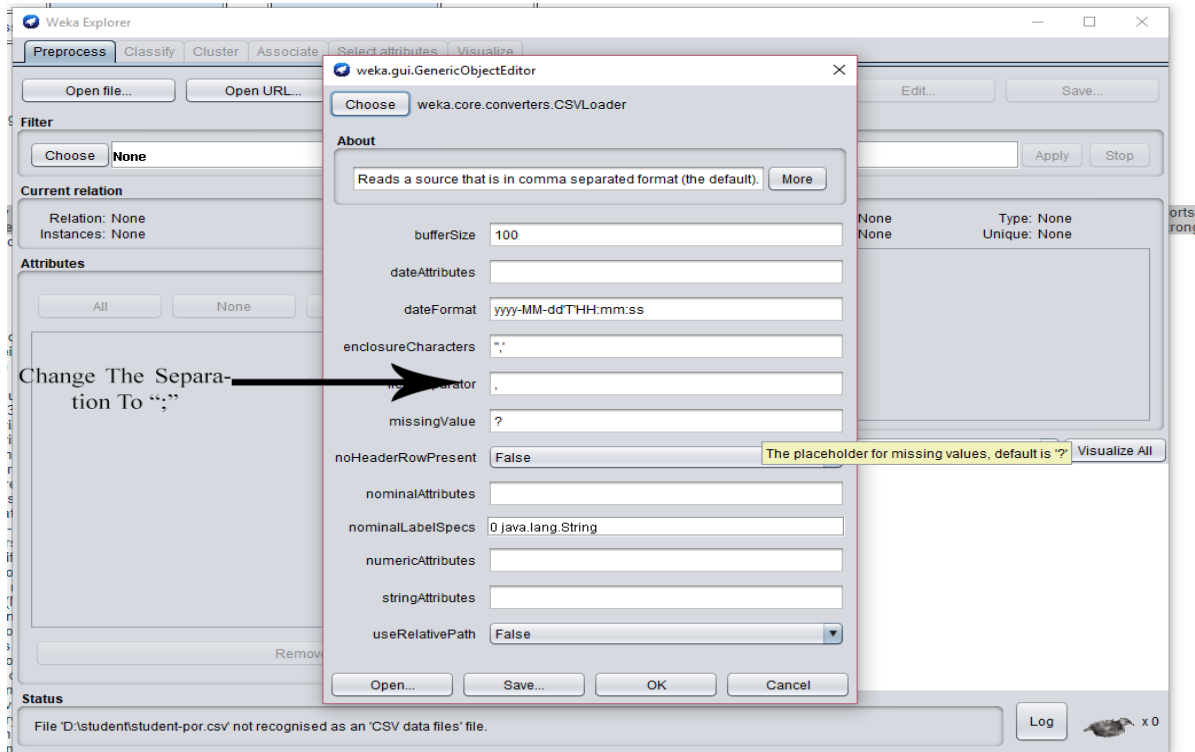
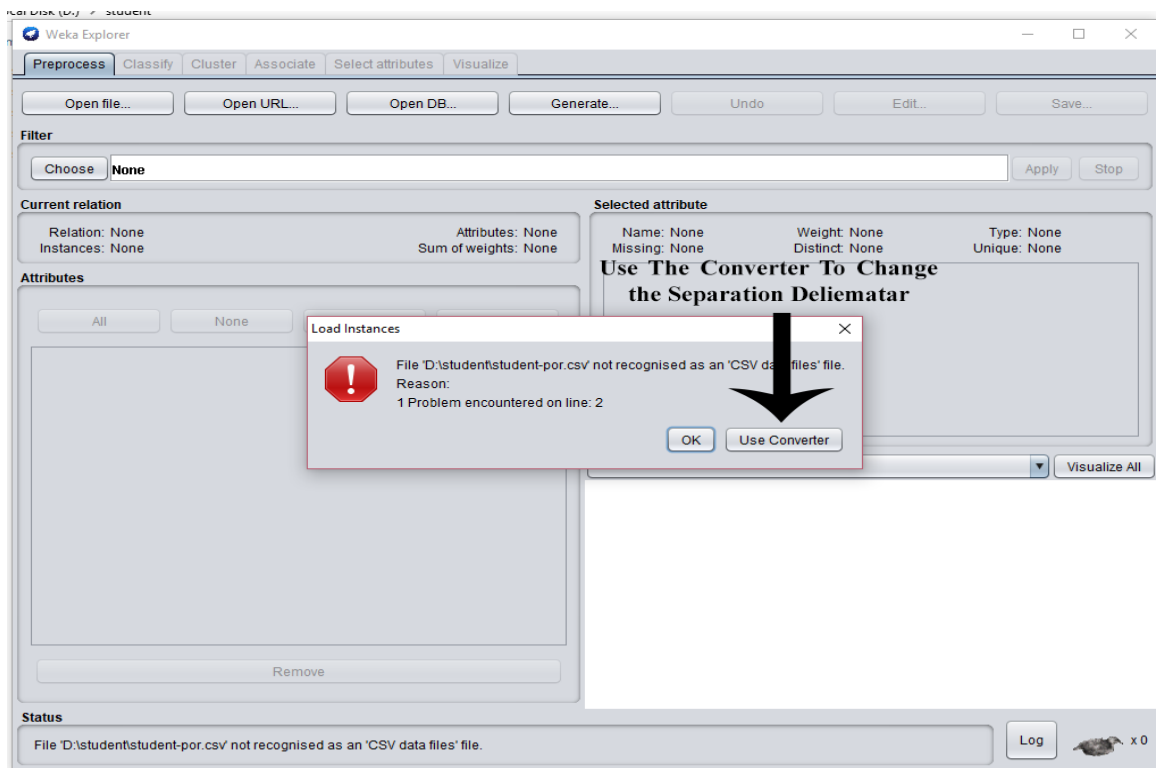
2-) Select Explore.



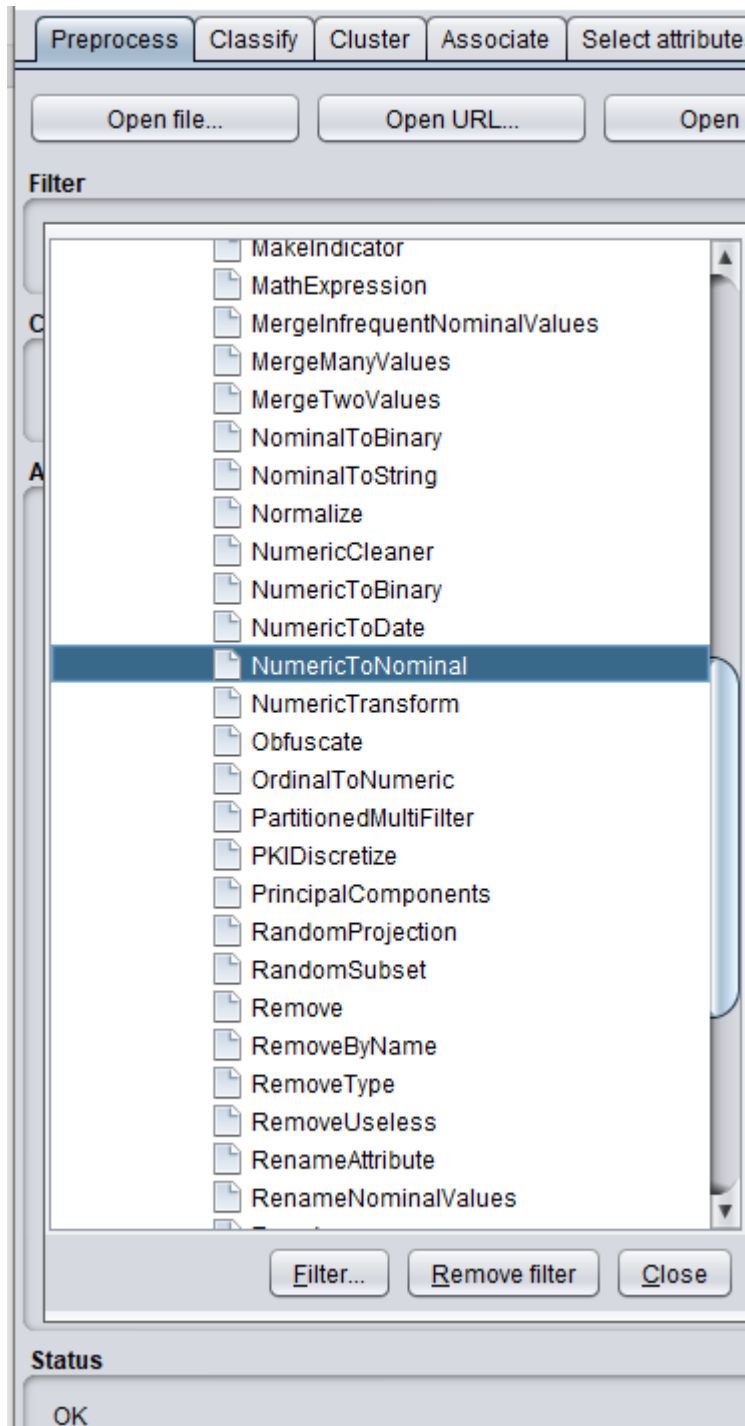
3-) Select Data Set.



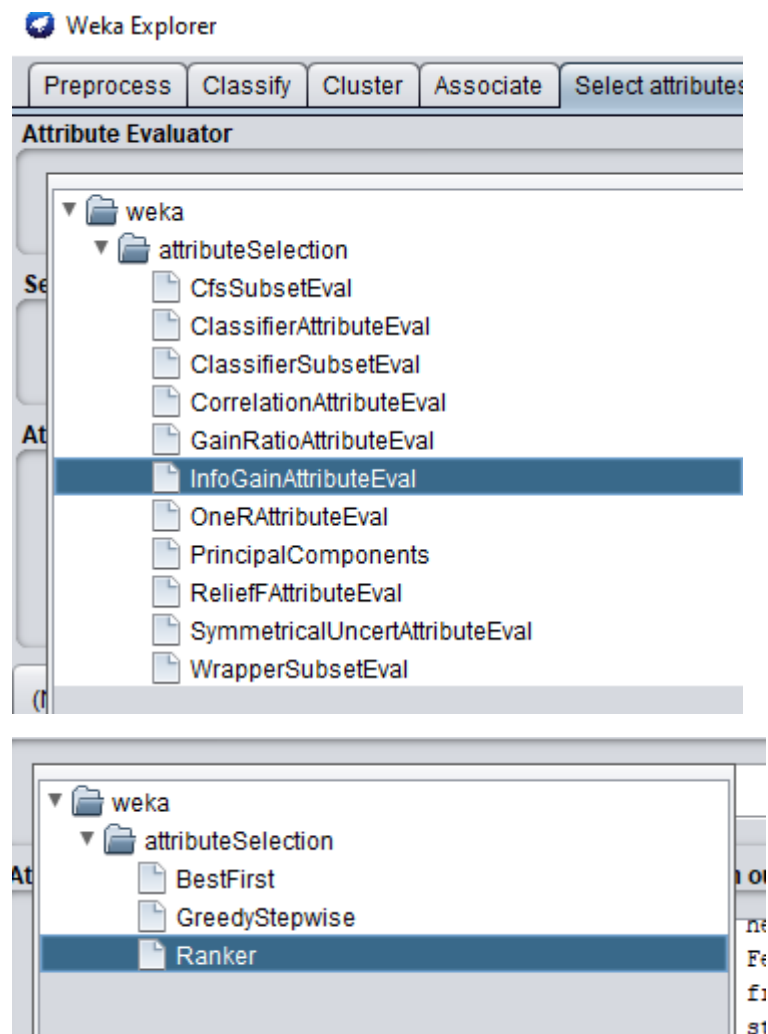
4-)Convert The Data Set and Select The Delimiter.



5-)Convert The Data to nominal.



6-)Getting Information Gain To Select Attributes.



output

```
healthn  
Fedu  
freetime  
studytime  
famrel  
reason  
traveltime  
schoolsup  
romantic  
paid  
guardian  
higher  
address  
sex  
internet  
nursery  
famsize  
activities  
Pstatus  
famsup  
school
```

```
tributes: 32,31,30,15,3,9,28,7,26,27,10,29,8,25,14,24,11,13,16,23,18,12,21,4,2,22,20,5,1
```

Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Open file...

Open URL...

Open DB...

Generate...

Undo

Edit...

Save...

Filter

Choose NumericToNominal - R first-last

Apply

Stop

Current relation

Relation: student-mat-weka.filters.unsupervised.attribut...
Instances: 395

Attributes: 6
Sum of weights: 395

Attributes

All

None

Invert

Pattern

No.	Name
1	<input checked="" type="checkbox"/> age
2	<input type="checkbox"/> failures
3	<input type="checkbox"/> absences
4	<input type="checkbox"/> G1
5	<input type="checkbox"/> G2
6	<input type="checkbox"/> G3

Remove

Selected attribute

Name: age
Missing: 0 (0%)
Distinct: 8
Type: Nominal
Unique: 2 (1%)

No.	Label	Count	Weight
1	15	82	82.0
2	16	104	104.0
3	17	98	98.0
4	18	82	82.0
5	19	24	24.0
6	20	3	3.0
7	21	1	1.0
8	22	1	1.0

Class: G3 (Nom)

Visualize All

Age	Count
15	82
16	104
17	98
18	82
19	24
20	3
21	1
22	1

Status

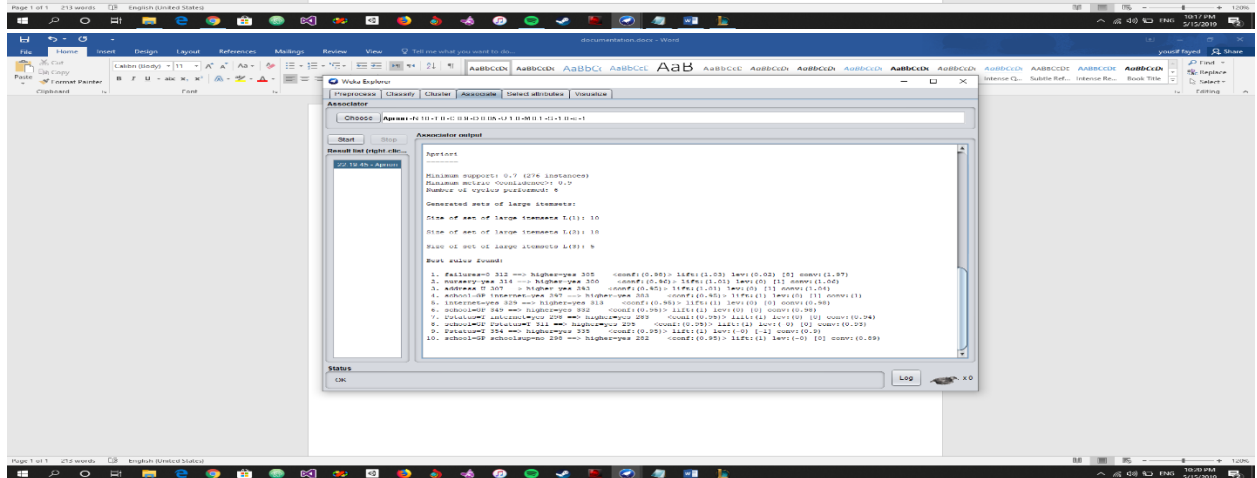
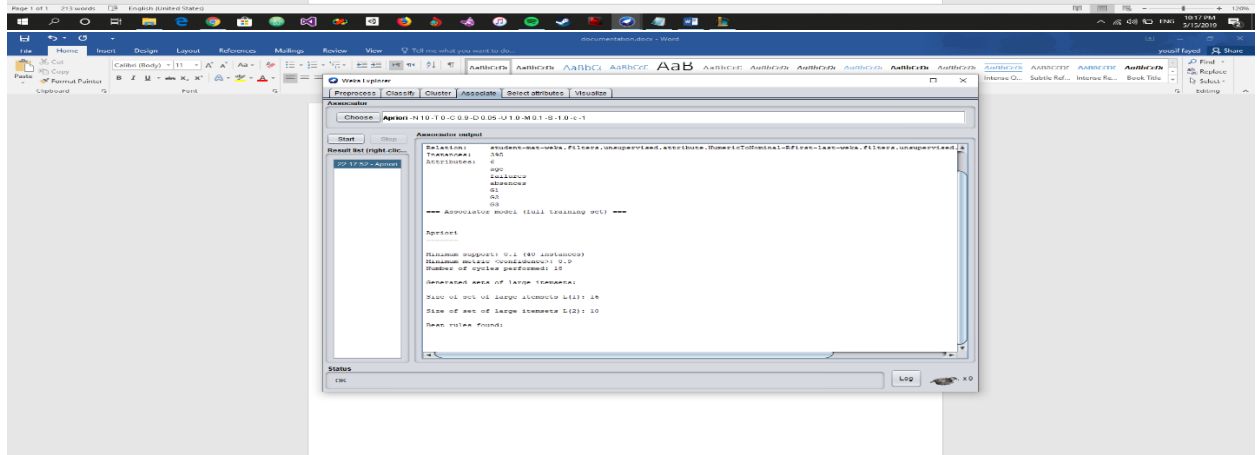
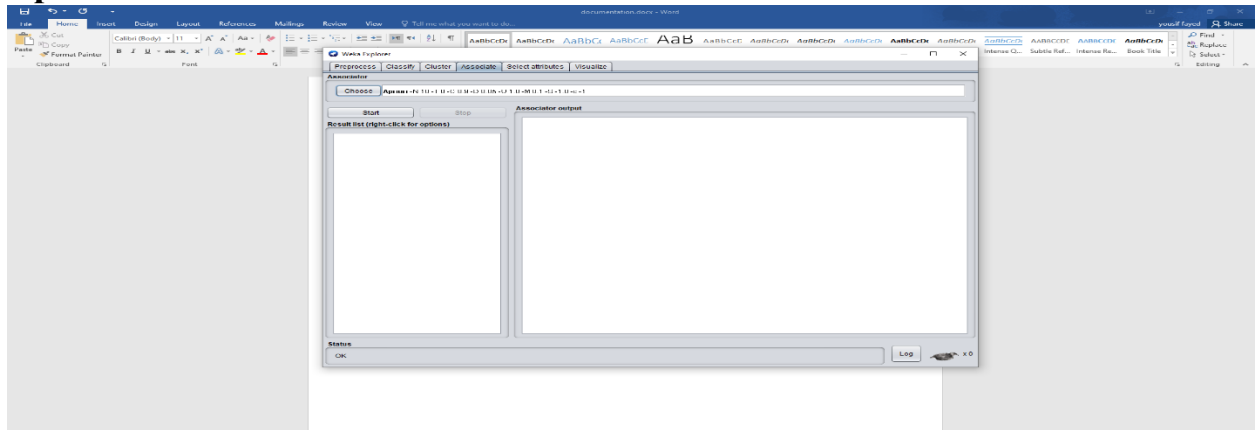
OK

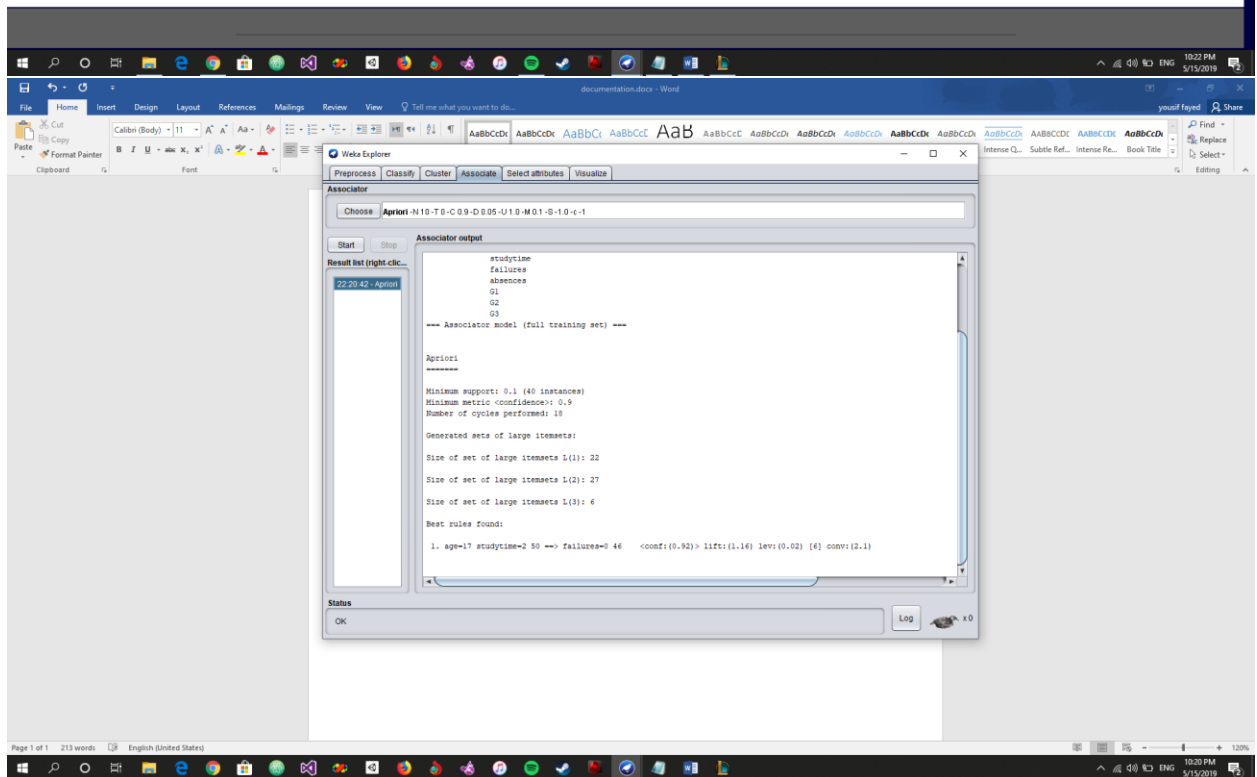
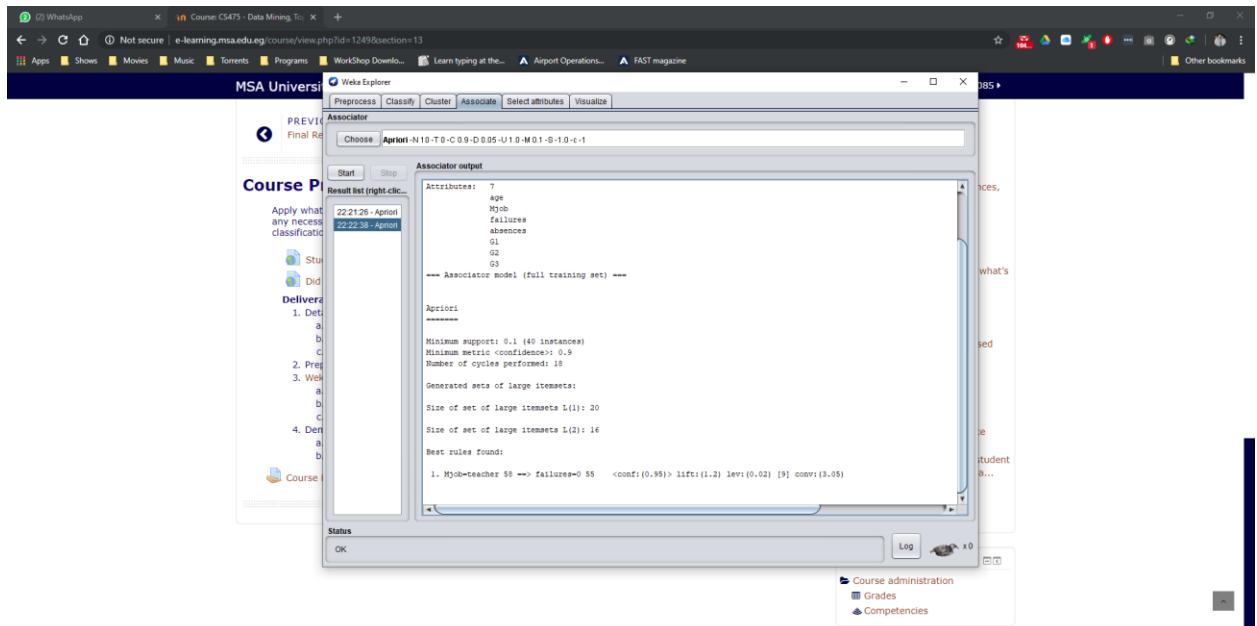
Log

x 0

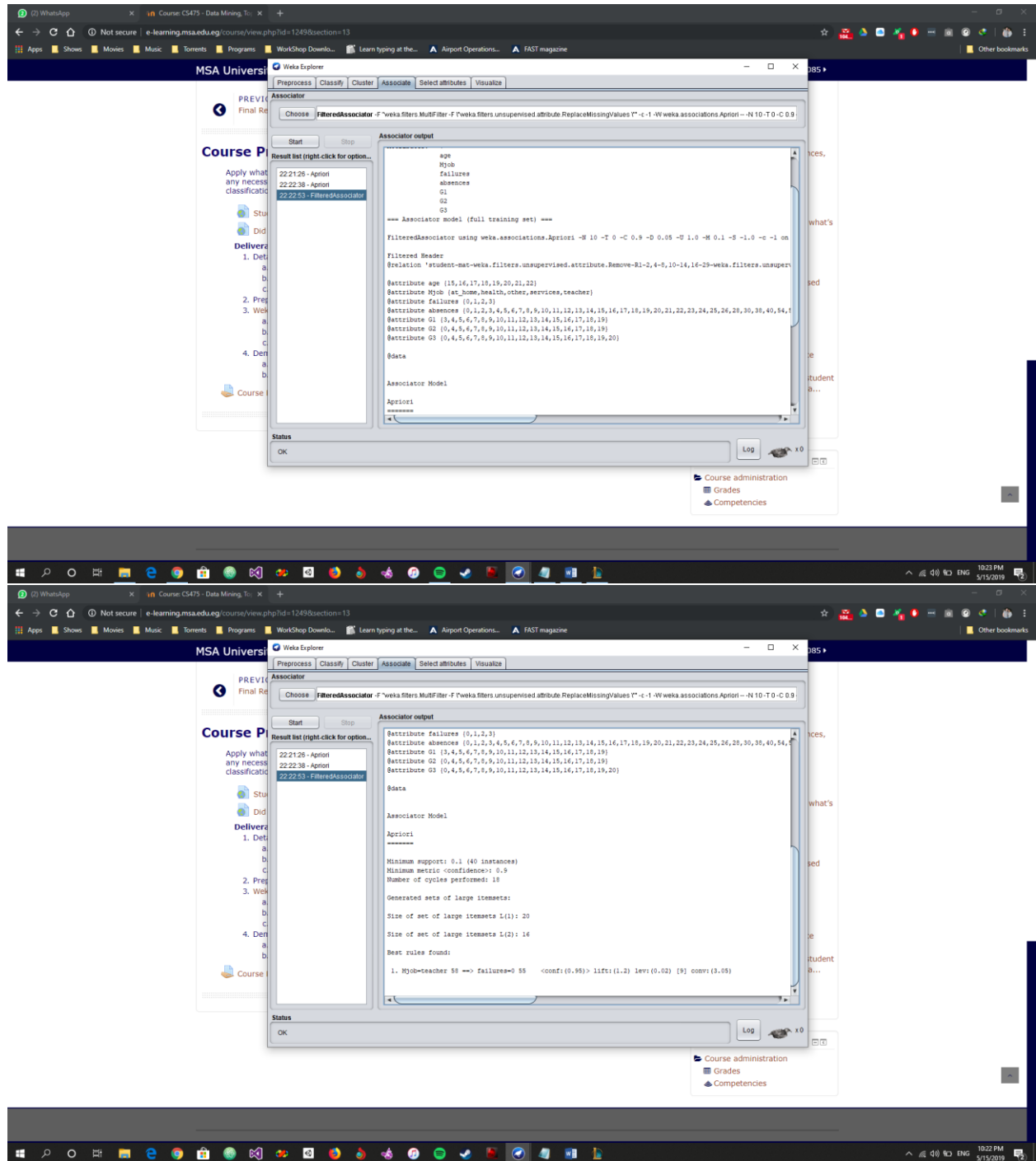
7-)Algorithms.

1) Apriori:



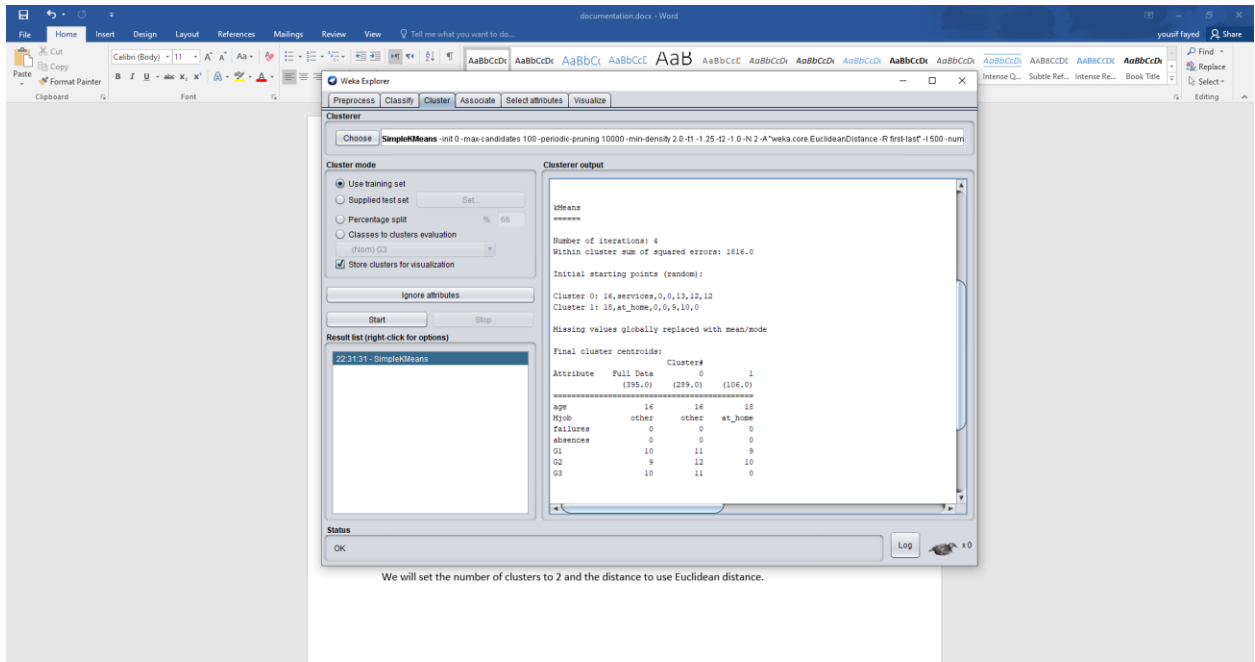


2) Filtered associator:



1. Clustering:

1.1.K-means:

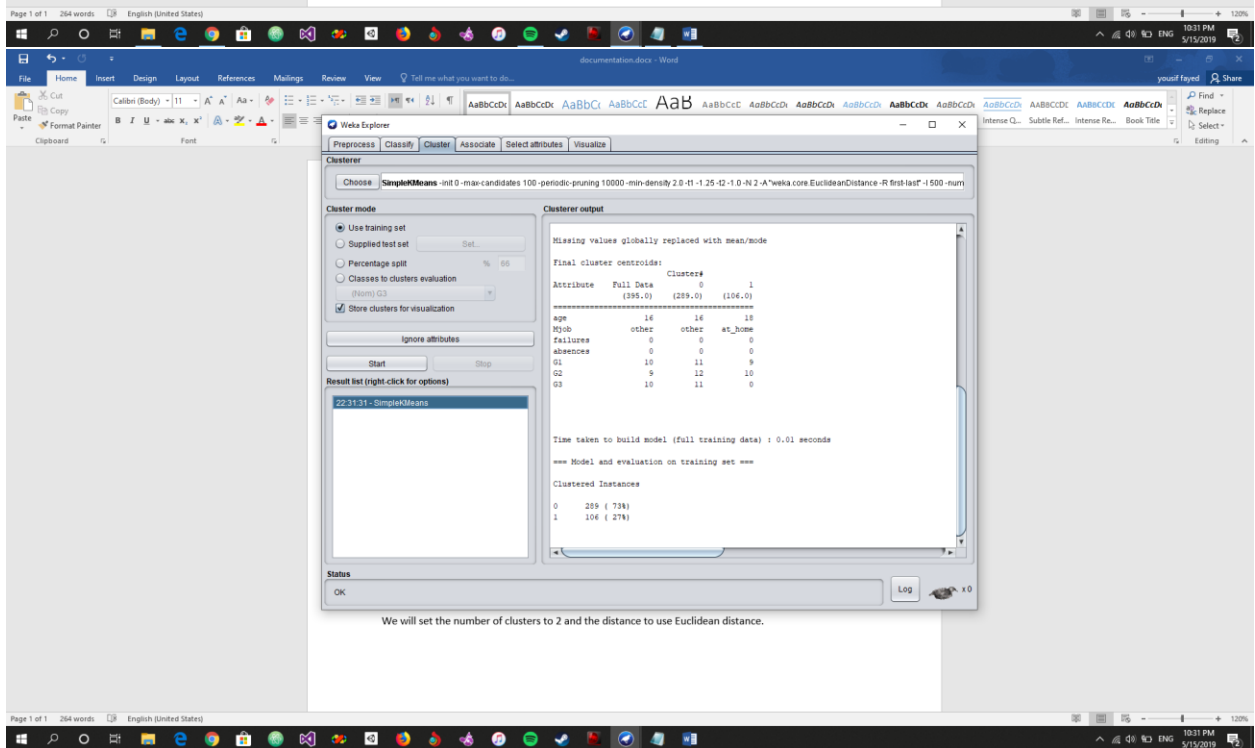


The screenshot shows the Weka Explorer interface with the 'Cluster' tab selected. The 'Clusterer' dropdown is set to 'SimpleKMeans'. The 'Cluster mode' section has 'Use training set' selected. The 'Cluster output' window displays the following results:

```
kmeans
=====
Number of iterations: 4
Within cluster sum of squared errors: 1016.0
Initial starting points (random):
Cluster 0: 16, services, 0, 0, 13, 12, 12
Cluster 1: 10, at_home, 0, 0, 9, 10, 0
Missing values globally replaced with mean/mode

Final cluster centroids:
Attribute      Full Data      Cluster#
(395.0)      (289.0)      (104.0)
=====
age            16            16      10
Mjob            other         other    at_home
failures       0             0        0
absences       0             0        0
G1             10            11       9
G2             9             12       10
G3            10            11       0
```

Below the 'Cluster output' window, the text reads: "We will set the number of clusters to 2 and the distance to use Euclidean distance."



The screenshot shows the Weka Explorer interface with the 'Cluster' tab selected. The 'Clusterer' dropdown is set to 'SimpleKMeans'. The 'Cluster mode' section has 'Use training set' selected. The 'Cluster output' window displays the following results:

```
Missing values globally replaced with mean/mode

Final cluster centroids:
Attribute      Full Data      Cluster#
(395.0)      (289.0)      (104.0)
=====
age            16            16      10
Mjob            other         other    at_home
failures       0             0        0
absences       0             0        0
G1             10            11       9
G2             9             12       10
G3            10            11       0

Time taken to build model (full training data) : 0.01 seconds

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

Clustered Instances
0      259 ( 73%)
1      106 ( 27%)
```

Below the 'Cluster output' window, the text reads: "We will set the number of clusters to 2 and the distance to use Euclidean distance."

1.2.EM:

The screenshot shows the Weka Explorer interface with the 'Clusterer' window open. The 'Choose' dropdown is set to 'EM-100-N2-K10-max-1-B-cv1-DE-6-M1-DE-6-K10-num-slots-1-S-100'. The 'Cluster mode' section has 'Use training set' selected. The 'Clusterer output' window displays the following data:

Cluster	Count	Percentage
0	32.9503	1.0017
9	32.9507	1.0042
10	33.9897	4.6113
11	31.7034	37.2366
12	2.048	30.535
13	1.0132	31.9507
14	1.0039	27.5961
15	1.0010	31.9507
16	1.0024	16.9976
17	2.0044	6.2584
18	1.0041	12.9409
19	2.0152	6.2584
20	1.0100	1.9998
[total]	322.156	225.514

Time taken to build model (full training data) : 0.00 seconds

Model and evaluation on training set

Clustered Instances

Cluster	Count	Percentage
0	106	50%
1	199	50%

Log likelihood: -12.02614

Status: OK

We will set the number of clusters to 2 and the distance to use Euclidean distance.

The second algorithm will be EM. The number of cluster can be set to 2 and by this the model will try to determine the number itself but we will set it to 2 to keep it consistent.

The screenshot shows the Weka Explorer interface with the 'Clusterer' window open. The 'Choose' dropdown is set to 'EM-100-N2-K10-max-1-B-cv1-DE-6-M1-DE-6-K10-num-slots-1-S-100'. The 'Cluster mode' section has 'Use training set' selected. The 'Clusterer output' window displays the following data:

Cluster	Count	Percentage
0	50.9415	1.0025
9	40.9424	7.0074
10	8.9791	29.0010
11	2.9497	41.0025
12	4.0019	29.9992
13	1.0014	16.9999
14	1.0031	13.9999
15	1.0047	8.9999
16	1.0152	1.9999
17	1.0152	1.9999
18	1.0152	1.9999
19	1.0152	1.9999
[total]	811.156	617.514

Time taken to build model (full training data) : 0.00 seconds

Model and evaluation on training set

Clustered Instances

Cluster	Count	Percentage
0	32.9503	1.0017
9	32.9507	1.0042
10	33.9897	4.6113
11	31.7034	37.2366
12	2.048	30.535
13	1.0132	31.9507
14	1.0039	27.5961
15	1.0010	31.9507
16	1.0024	16.9976
17	2.0044	6.2584
18	1.0041	12.9409
19	2.0152	6.2584
20	1.0100	1.9998
[total]	322.156	225.514

Log likelihood: -12.02614

Status: OK

We will set the number of clusters to 2 and the distance to use Euclidean distance.

The second algorithm will be EM. The number of cluster can be set to 2 and by this the model will try to determine the number itself but we will set it to 2 to keep it consistent.

The screenshot shows the Weka Explorer interface with the 'Clusterer' window open. The 'Choose' dropdown is set to 'EM-100-N2-K10-max-1-B-cv1-DE-6-M1-DE-6-K10-num-slots-1-S-100'. The 'Cluster mode' section has 'Use training set' selected. The 'Clusterer output' window displays the following data:

Cluster	Count	Percentage
0	10.49	0.31
1	19.51	0.69
[total]	30.00	1.00

Time taken to build model (full training data) : 0.00 seconds

Model and evaluation on training set

Clustered Instances

Cluster	Count	Percentage
0	10.49	0.31
1	19.51	0.69
[total]	30.00	1.00

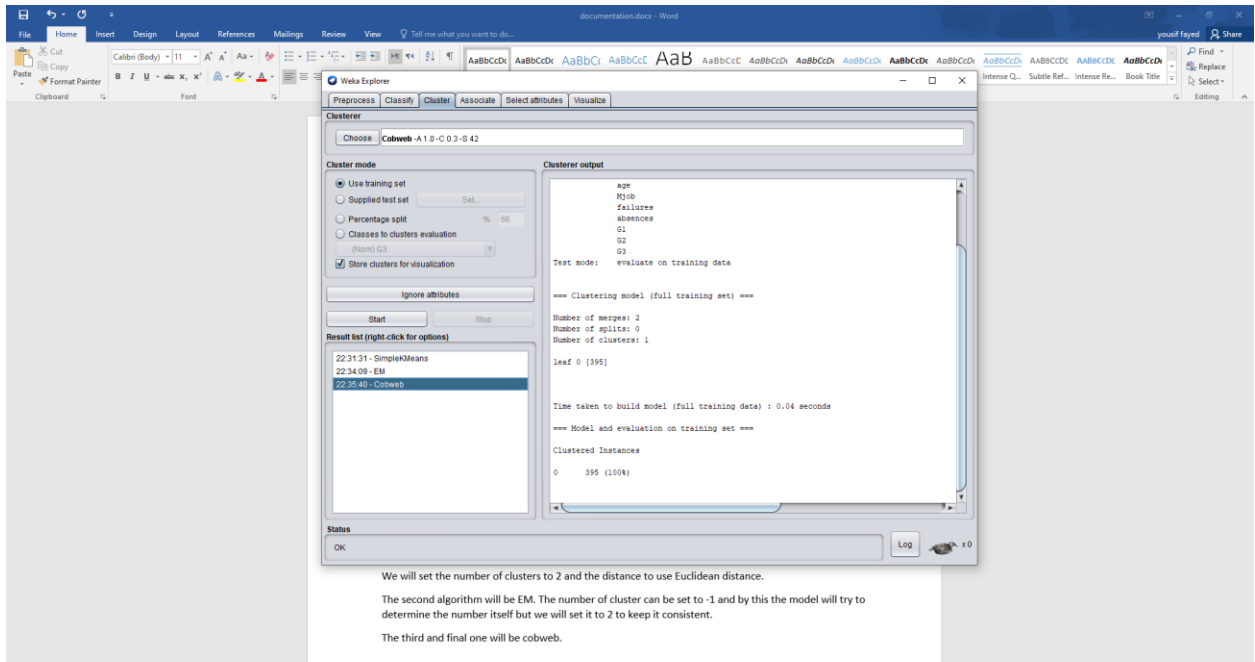
Log likelihood: -12.02614

Status: OK

We will set the number of clusters to 2 and the distance to use Euclidean distance.

The second algorithm will be EM. The number of cluster can be set to 2 and by this the model will try to determine the number itself but we will set it to 2 to keep it consistent.

1.3.Cobweb:



The screenshot shows the Weka Explorer interface with the 'Clusterer' window open. The 'Cluster mode' section has 'Use training set' selected. The 'Clusterer output' pane displays the following information:

```
age
Hjob
failures
absences
G1
G2
G3

Test mode: evaluate on training data

=== Clustering model (full training set) ===
Number of merges: 2
Number of splits: 0
Number of clusters: 1
leaf 0 [395]

Time taken to build model (full training data) : 0.04 seconds

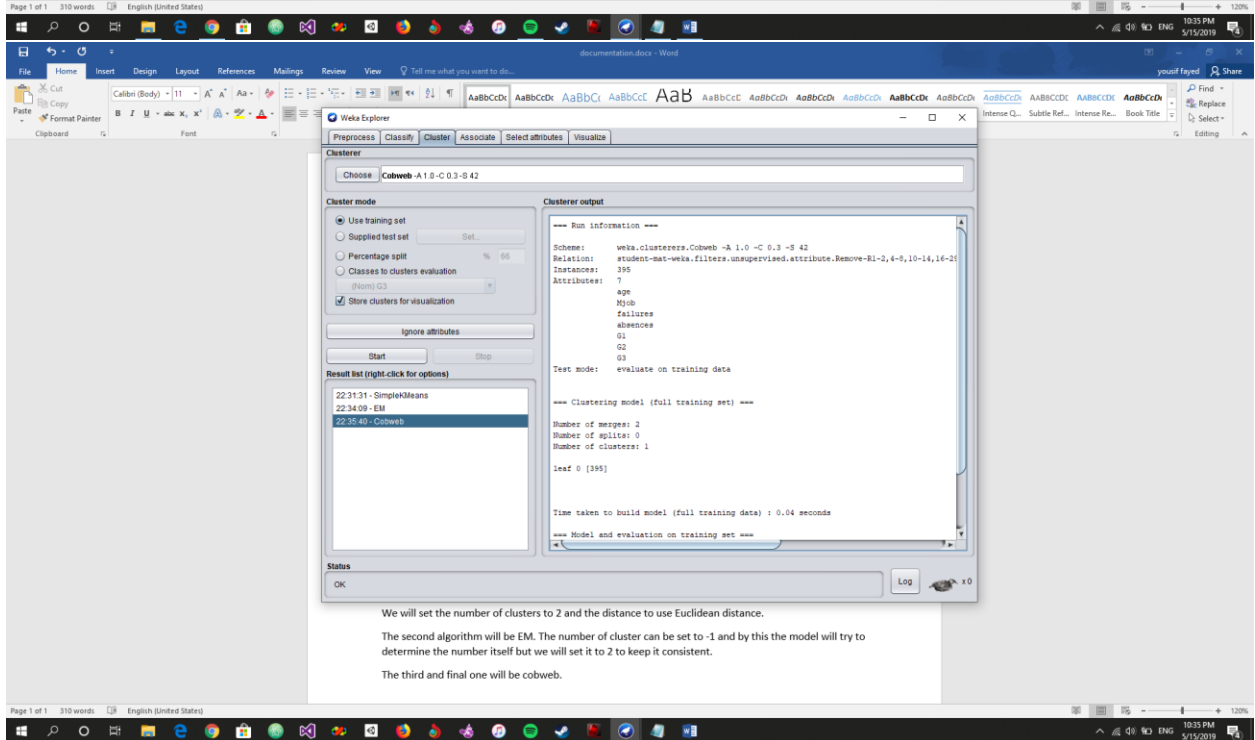
=== Model and evaluation on training set ===
Clustered Instances
0 395 (100%)
```

Below the window, the following text is present:

We will set the number of clusters to 2 and the distance to use Euclidean distance.

The second algorithm will be EM. The number of cluster can be set to -1 and by this the model will try to determine the number itself but we will set it to 2 to keep it consistent.

The third and final one will be cobweb.



The screenshot shows the Weka Explorer interface with the 'Clusterer' window open. The 'Cluster mode' section has 'Use training set' selected. The 'Clusterer output' pane displays the following information:

```
=== Run information ===
Scheme: weka.clusterers.Cobweb -A 1.0 -C 0.3 -S 42
Relation: student-mat-weka.filters.unsupervised.attribute.Remove-R1-2,4-5,10-14,16-20
Instances: 395
Attributes: 7
age
Hjob
failures
absences
G1
G2
G3

Test mode: evaluate on training data

=== Clustering model (full training set) ===
Number of merges: 2
Number of splits: 0
Number of clusters: 1
leaf 0 [395]

Time taken to build model (full training data) : 0.04 seconds

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

Below the window, the following text is present:

We will set the number of clusters to 2 and the distance to use Euclidean distance.

The second algorithm will be EM. The number of cluster can be set to -1 and by this the model will try to determine the number itself but we will set it to 2 to keep it consistent.

The third and final one will be cobweb.

- Classifications:
 - One-R:

Weka Explorer

Preprocess | **Classify** | Cluster | Associate | Select attributes | Visualize

Classifier

Choose OneR - B 6

Test options

☐ Use training set
☐ Supplied test set Set...
☒ Cross-validation Folds 10
☐ Percentage split % 66
 More options...

(Nom) G3

Start Stop

Result list (right-click for options)

22:43:48 - rules.OneR

Classifier output

```

      0.500  0.005  0.600  0.500  0.545  0.541  0.747  0.308  17
      0.750  0.008  0.750  0.750  0.750  0.742  0.871  0.570  18
      0.400  0.003  0.667  0.400  0.500  0.512  0.699  0.274  19
      0.000  0.000  ?      0.000  ?      ?      0.500  0.003  20
Weighted Avg.  0.501  0.052  ?      0.501  ?      ?      0.725  0.306

=== Confusion Matrix ===
 a b c d e f g h i j k l m n o p q r <-- classified as
18 0 2 3 0 8 3 4 0 0 0 0 0 0 0 0 0 0 | a = 0
 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 | b = 4
 5 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 | c = 5
 4 0 0 6 0 5 0 0 0 0 0 0 0 0 0 0 0 0 | d = 6
 0 0 0 2 0 6 0 1 0 0 0 0 0 0 0 0 0 0 | e = 7
 0 0 0 1 0 20 2 9 0 0 0 0 0 0 0 0 0 0 | f = 8
 0 0 0 0 0 7 1 19 1 0 0 0 0 0 0 0 0 0 | g = 9
 0 0 0 0 0 4 3 40 8 1 0 0 0 0 0 0 0 0 | h = 10
 0 0 0 0 0 2 0 12 19 13 1 0 0 0 0 0 0 0 | i = 11
 0 0 0 0 0 0 0 2 6 17 6 0 0 0 0 0 0 0 | j = 12
 0 0 0 0 0 0 0 0 1 8 19 3 0 0 0 0 0 0 | k = 13
 0 0 0 0 0 0 0 0 0 2 10 14 1 0 0 0 0 0 | l = 14
 0 0 0 0 0 0 0 0 0 0 1 6 23 3 0 0 0 0 | m = 15
 0 0 0 0 0 0 0 0 0 0 0 0 9 7 0 0 0 0 | n = 16
 0 0 0 0 0 0 0 0 0 0 0 0 0 1 2 3 0 0 | o = 17
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 2 9 0 | p = 18
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 3 2 0 | q = 19
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 | r = 20
  
```

Status

OK Log x0

Weka Explorer

Preprocess | **Classify** | Cluster | Associate | Select attributes | Visualize

Classifier

Choose OneR - B 6

Test options

☐ Use training set
☐ Supplied test set Set...
☒ Cross-validation Folds 10
☐ Percentage split % 66
 More options...

(Nom) G3

Start Stop

Result list (right-click for options)

22:43:48 - rules.OneR

Classifier output

```

=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances      198          50.1266 %
Incorrectly Classified Instances    197          49.8734 %
Kappa statistic                    0.451
Mean absolute error                 0.0554
Root mean squared error             0.2354
Relative absolute error             54.3707 %
Root relative squared error         104.3406 %
Total Number of Instances          395

=== Detailed Accuracy By Class ===
      TP Rate  FP Rate  Precision  Recall  F-Measure  MCC  ROC Area  PRC Area  Cla
0.474  0.025  0.667  0.474  0.554  0.524  0.724  0.366  0
0.000  0.000  ?      0.000  ?      ?      0.500  0.003  4
0.000  0.005  0.000  0.000  0.000  -0.010  0.497  0.018  5
0.400  0.021  0.429  0.400  0.414  0.392  0.689  0.194  6
0.000  0.000  ?      0.000  ?      ?      0.500  0.023  7
0.625  0.091  0.377  0.625  0.471  0.428  0.767  0.266  8
0.036  0.022  0.111  0.036  0.054  0.024  0.507  0.072  9
0.714  0.139  0.460  0.714  0.559  0.485  0.788  0.369  10
0.404  0.046  0.543  0.404  0.463  0.408  0.679  0.290  11
0.548  0.066  0.415  0.548  0.472  0.425  0.741  0.263  12
0.613  0.049  0.514  0.613  0.559  0.520  0.782  0.345  13
0.519  0.024  0.609  0.519  0.560  0.532  0.747  0.349  14
0.697  0.030  0.676  0.697  0.687  0.658  0.833  0.497  15
  
```

Status

OK Log x0

○ **J48:**

Test options

- ☐ Use training set
- ☐ Supplied test set
- ☒ Cross-validation Folds
- ☐ Percentage split %

(Nom) G3

Status
OK Log

```
Preprocess | Classify | Cluster | Associate | Select attributes | Visualize
```

Classifier

J48 - C 0.25 - M 2

Test options

- ☐ Use training set
- ☐ Supplied test set
- ☒ Cross-validation Folds
- ☐ Percentage split %

(Nom) G3

Result list (right-click for options)

22:43:48 - rules.OneR
22:45:03 - trees.J48

Classifier output

```
Kappa statistic          0.438
Mean absolute error      0.0696
Root mean squared error   0.1987
Relative absolute error    68.2931 %
Root relative squared error 88.0776 %
Total Number of Instances 395

=== Detailed Accuracy By Class ===
```

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.526	0.034	0.625	0.526	0.571	0.532	0.849	0.638	0	
0.000	0.003	0.000	0.000	0.000	-0.003	0.487	0.003	4	
0.286	0.003	0.667	0.286	0.400	0.430	0.842	0.379	5	
0.400	0.021	0.429	0.400	0.414	0.392	0.826	0.266	6	
0.000	0.003	0.000	0.000	0.000	-0.008	0.860	0.109	7	
0.625	0.091	0.377	0.625	0.471	0.428	0.856	0.293	8	
0.107	0.030	0.214	0.107	0.143	0.107	0.742	0.176	9	
0.607	0.118	0.459	0.607	0.523	0.437	0.857	0.401	10	
0.426	0.072	0.444	0.426	0.435	0.360	0.854	0.367	11	
0.355	0.052	0.367	0.355	0.361	0.307	0.839	0.272	12	
0.613	0.049	0.514	0.613	0.559	0.520	0.897	0.375	13	
0.519	0.027	0.583	0.519	0.549	0.519	0.900	0.388	14	
0.697	0.030	0.676	0.697	0.687	0.658	0.941	0.525	15	
0.438	0.016	0.538	0.438	0.483	0.466	0.957	0.365	16	
0.500	0.005	0.600	0.500	0.545	0.541	0.894	0.388	17	
0.750	0.008	0.750	0.750	0.750	0.742	0.948	0.594	18	
0.400	0.003	0.667	0.400	0.500	0.512	0.985	0.448	19	
0.000	0.000	?	0.000	?	?	0.496	0.003	20	
Weighted Avg.	0.489	0.052	?	0.489	?	?	0.865	0.383	


```
=== Confusion Matrix ===
```

a b c d e f g h i j k l m n o p q r <-- classified as
20 1 0 2 0 8 3 4 0 0 0 0 0 0 0 0 0 0 a = 0
0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 b = 4
3 0 2 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 c = 5
4 0 1 6 1 3 0 0 0 0 0 0 0 0 0 0 0 0 d = 6
0 0 0 2 0 6 0 1 0 0 0 0 0 0 0 0 0 0 e = 7
2 0 0 1 0 20 1 8 0 0 0 0 0 0 0 0 0 0 f = 8
1 0 0 1 0 8 3 14 1 0 0 0 0 0 0 0 0 0 g = 9
2 0 0 0 0 5 6 34 9 0 0 0 0 0 0 0 0 0 h = 10
0 0 0 0 0 2 1 11 20 12 1 0 0 0 0 0 0 0 i = 11
0 0 0 0 0 0 0 0 2 11 11 6 1 0 0 0 0 0 j = 12
0 0 0 0 0 0 0 0 4 5 19 3 0 0 0 0 0 0 k = 13
0 0 0 0 0 0 0 0 0 2 10 14 1 0 0 0 0 0 l = 14
0 0 0 0 0 0 0 0 0 0 0 1 6 23 3 0 0 0 0 m = 15
0 0 0 0 0 0 0 0 0 0 0 0 9 7 0 0 0 0 n = 16
0 0 0 0 0 0 0 0 0 0 0 0 0 1 2 3 0 0 0 o = 17
0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 2 9 0 0 p = 18
0 0 0 0 0 0 0 0 0 0 0 0 0 0 3 2 0 0 q = 19
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 r = 20

○ Naïve Bayes:

Preprocess **Classify** Cluster Associate Select attributes Visualize

Classifier: Choose **NaiveBayes**

Test options

☐ Use training set

☐ Supplied test set

☒ Cross-validation Folds

☐ Percentage split %

(Nom) G3

Result list (right-click for options)

- 22:43:48 - rules.OneR
- 22:45:03 - trees.J48
- 22:46:00 - bayes.NaiveBayes

Classifier output

	0.000	0.000	?	0.000	?	?	0.959	0.223	17
	0.750	0.010	0.692	0.750	0.720	0.711	0.966	0.699	18
	0.200	0.000	1.000	0.200	0.333	0.445	0.942	0.640	19
	0.000	0.000	?	0.000	?	?	0.876	0.020	20
Weighted Avg.	0.451	0.059	?	0.451	?	?	0.904	0.468	

=== Confusion Matrix ===

a b c d e f g h i j k l m n o p q r <-- classified as

29 0 0 0 0 3 1 5 0 0 0 0 0 0 0 0 0 0 0 | a = 0

1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 | b = 4

3 0 0 2 0 2 0 0 0 0 0 0 0 0 0 0 0 0 | c = 5

2 0 0 9 0 1 1 2 0 0 0 0 0 0 0 0 0 0 | d = 6

3 0 0 2 0 2 2 0 0 0 0 0 0 0 0 0 0 0 | e = 7

5 0 1 1 0 8 7 10 0 0 0 0 0 0 0 0 0 0 | f = 8

0 0 0 0 0 5 4 18 1 0 0 0 0 0 0 0 0 0 | g = 9

2 0 0 0 0 3 1 41 9 0 0 0 0 0 0 0 0 0 | h = 10

0 0 0 0 0 1 0 14 21 8 3 0 0 0 0 0 0 0 | i = 11

0 0 0 0 0 0 0 2 11 9 7 2 0 0 0 0 0 0 | j = 12

0 0 0 0 0 0 0 0 4 5 10 12 0 0 0 0 0 0 | k = 13

0 0 0 0 0 0 0 0 2 2 7 14 2 0 0 0 0 0 | l = 14

0 0 0 0 0 0 0 0 1 0 0 6 23 3 0 0 0 0 | m = 15

1 0 0 0 0 0 0 0 0 0 0 0 15 0 0 0 0 0 | n = 16

0 0 0 0 0 0 0 0 0 0 0 0 0 4 1 0 1 0 0 | o = 17

0 0 0 0 0 0 0 0 0 0 0 0 0 3 0 9 0 0 | p = 18

1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 3 1 0 0 | q = 19

0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 | r = 20

Status: OK x0

○ Logistic:

Weka Explorer

Preprocess **Classify** Cluster Associate Select attributes Visualize

Classifier: Choose **Logistic -R 1.0E-8 -M -1 -num-decimal-places 4**

Test options

☒ Use training set

☐ Supplied test set

☐ Cross-validation Folds

☐ Percentage split %

(Nom) G3

Result list (right-click for options)

- 22:43:48 - rules.OneR
- 22:45:03 - trees.J48
- 22:46:00 - bayes.NaiveBayes
- 22:47:57 - functions.Logistic

Classifier output

	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	17
	1.000	0.003	0.923	1.000	0.960	0.960	1.000	0.994	18
	0.800	0.000	1.000	0.800	0.889	0.893	1.000	0.967	19
	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	20
Weighted Avg.	0.881	0.013	0.882	0.881	0.880	0.868	0.992	0.946	

=== Confusion Matrix ===

a b c d e f g h i j k l m n o p q r <-- classified as

37 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 | a = 0

0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 | b = 4

0 0 7 0 0 0 0 0 0 0 0 0 0 0 0 0 0 | c = 5

0 0 0 15 0 0 0 0 0 0 0 0 0 0 0 0 0 | d = 6

0 0 0 0 9 0 0 0 0 0 0 0 0 0 0 0 0 | e = 7

0 0 0 0 0 31 0 1 0 0 0 0 0 0 0 0 0 | f = 8

1 0 0 0 0 0 22 4 1 0 0 0 0 0 0 0 0 | g = 9

0 0 0 0 0 1 3 49 2 1 0 0 0 0 0 0 0 | h = 10

1 0 0 0 0 0 1 5 35 5 0 0 0 0 0 0 0 | i = 11

0 0 0 0 0 0 1 2 3 21 3 1 0 0 0 0 0 | j = 12

0 0 0 0 0 0 0 0 0 2 27 2 0 0 0 0 0 | k = 13

0 0 0 0 0 0 0 0 1 2 2 22 0 0 0 0 0 | l = 14

0 0 0 0 0 0 0 0 0 0 0 0 33 0 0 0 0 | m = 15

0 0 0 0 0 0 0 0 0 0 0 0 0 16 0 0 0 | n = 16

0 0 0 0 0 0 0 0 0 0 0 0 0 0 6 0 0 0 | o = 17

0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 12 0 0 | p = 18

0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 4 0 0 | q = 19

0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 | r = 20

Status: OK x0