DATA MINING LAB

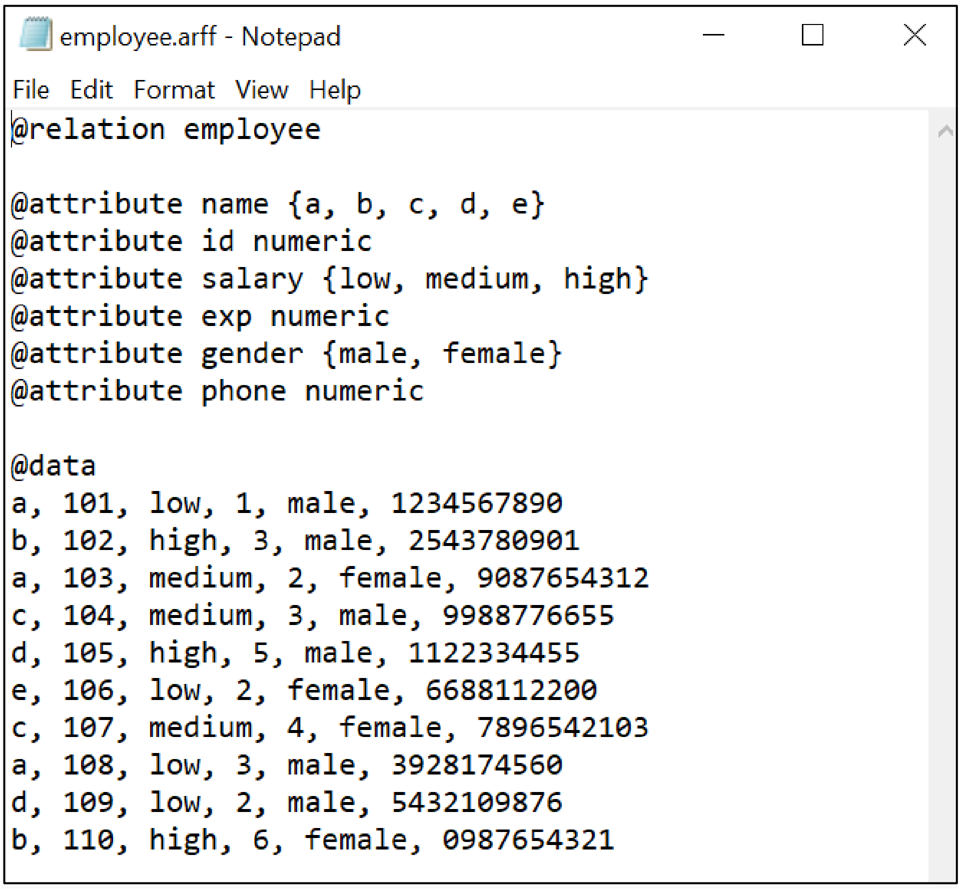
ASSIGNMENT-1.2

Ayush Kumar

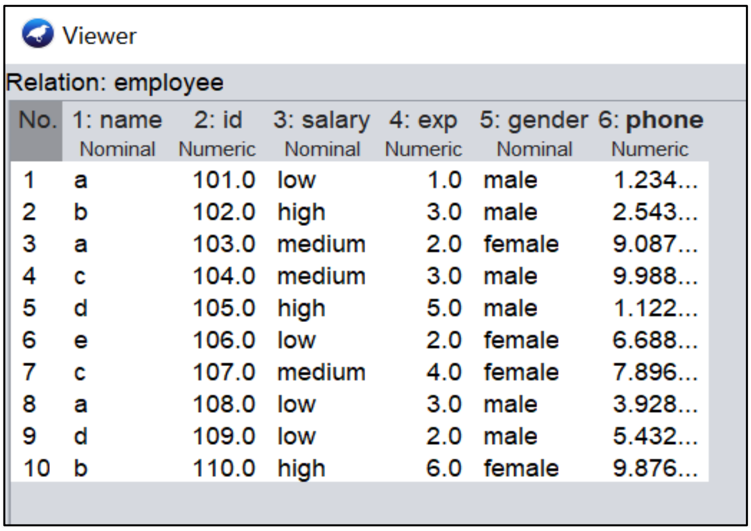
20214284

CSE-6B1

1) Create an Employee Table with training data set which includes attributes like name, id, salary, experience, gender, phone number with the help of Data Mining Tool WEKA.

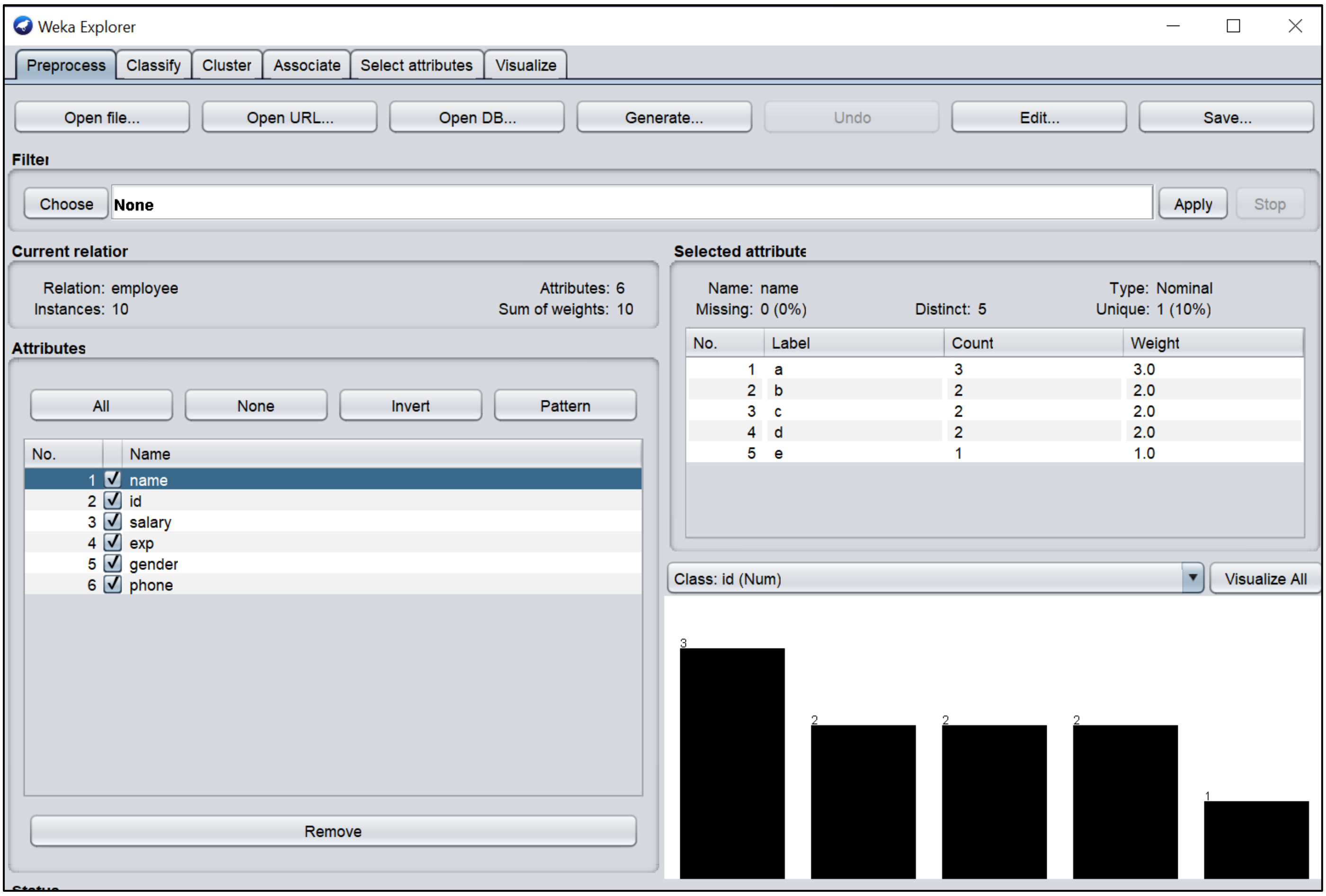


```
@relation employee

@attribute name {a, b, c, d, e}
@attribute id numeric
@attribute salary {low, medium, high}
@attribute exp numeric
@attribute gender {male, female}
@attribute phone numeric

@data
a, 101, low, 1, male, 1234567890
b, 102, high, 3, male, 2543780901
a, 103, medium, 2, female, 9087654312
c, 104, medium, 3, male, 9988776655
d, 105, high, 5, male, 1122334455
e, 106, low, 2, female, 6688112200
c, 107, medium, 4, female, 7896542103
a, 108, low, 3, male, 3928174560
d, 109, low, 2, male, 5432109876
b, 110, high, 6, female, 0987654321
```



Relation: employee

No.	1: name	2: id	3: salary	4: exp	5: gender	6: **phone**
	Nominal	Numeric	Nominal	Numeric	Nominal	Numeric
1	a	101.0	low	1.0	male	1.234...
2	b	102.0	high	3.0	male	2.543...
3	a	103.0	medium	2.0	female	9.087...
4	c	104.0	medium	3.0	male	9.988...
5	d	105.0	high	5.0	male	1.122...
6	e	106.0	low	2.0	female	6.688...
7	c	107.0	medium	4.0	female	7.896...
8	a	108.0	low	3.0	male	3.928...
9	d	109.0	low	2.0	male	5.432...
10	b	110.0	high	6.0	female	9.876...



Filter: None

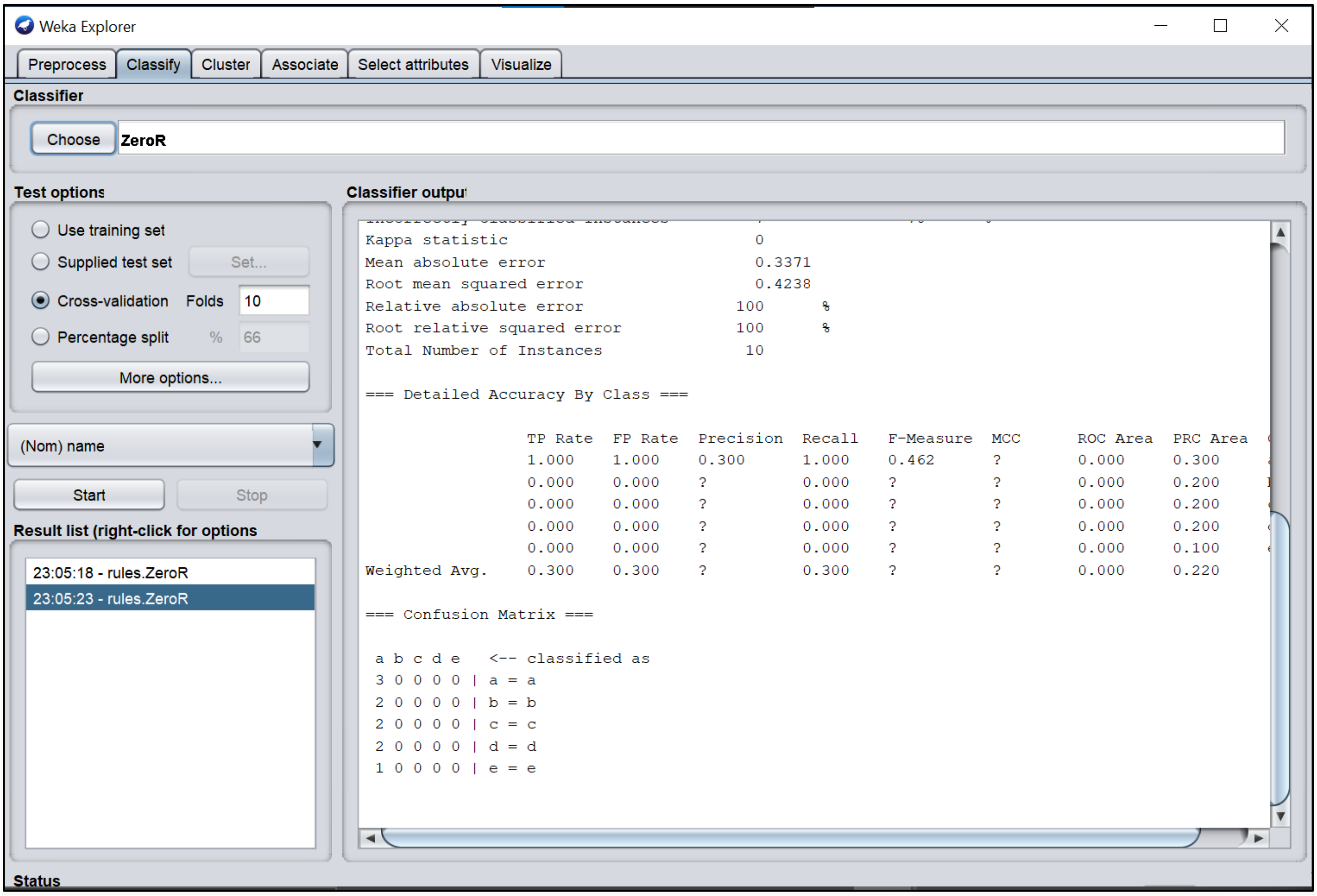
Relation: employee Attributes: 6
Instances: 10 Sum of weights: 10

No.		Name
1	☑	name
2	☑	id
3	☑	salary
4	☑	exp
5	☑	gender
6	☑	phone

Name: name Type: Nominal
Missing: 0 (0%) Distinct: 5 Unique: 1 (10%)

No.	Label	Count	Weight
1	a	3	3.0
2	b	2	2.0
3	c	2	2.0
4	d	2	2.0
5	e	1	1.0

Class: id (Num)



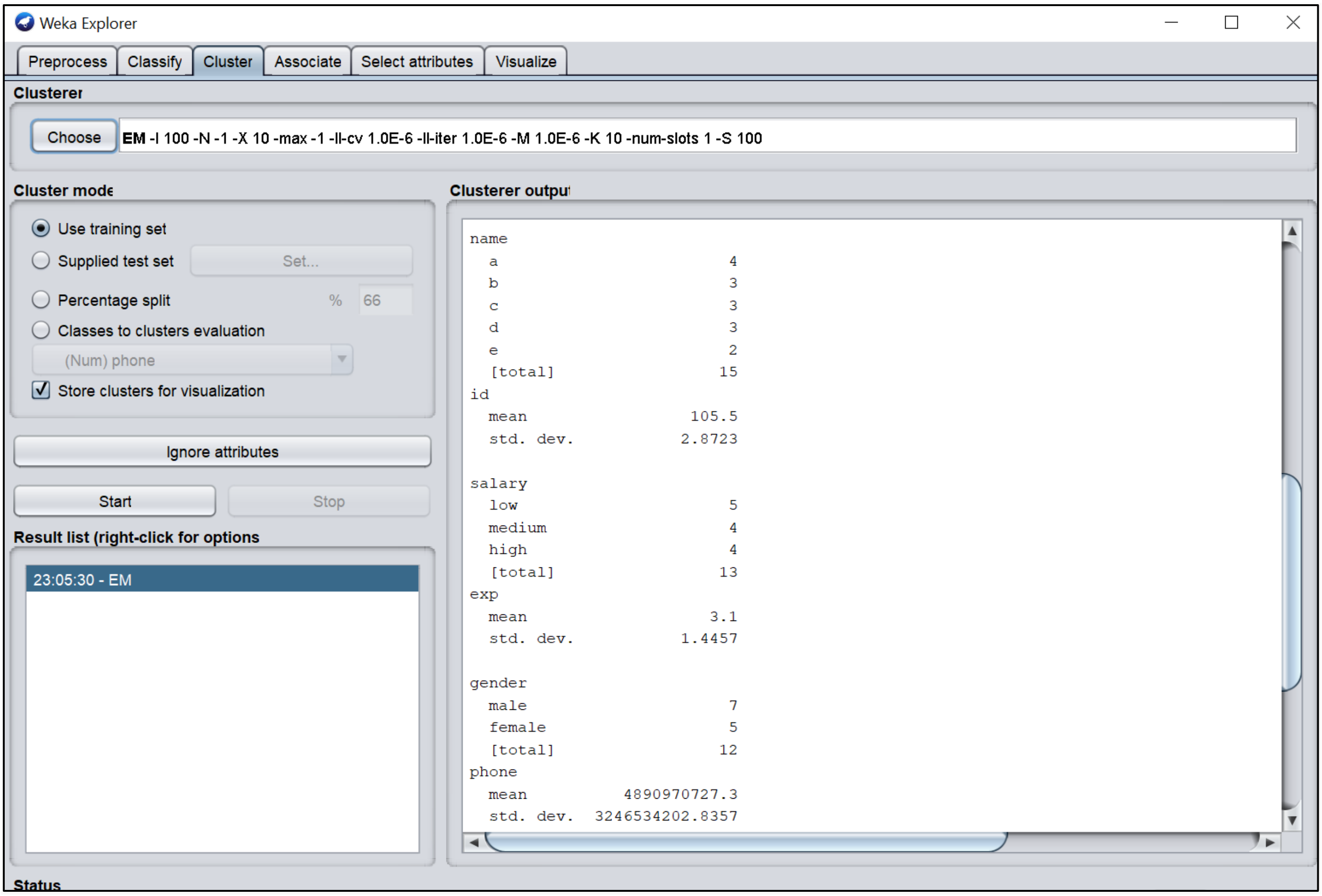
Classifier: ZeroR

Test options:
○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) name

Result list:
23:05:18 - rules.ZeroR
23:05:23 - rules.ZeroR

```
Kappa statistic                          0
Mean absolute error                      0.3371
Root mean squared error                  0.4238
Relative absolute error                100      %
Root relative squared error            100      %
Total Number of Instances               10

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area
                 1.000    1.000    0.300      1.000    0.462      ?        0.000     0.300
                 0.000    0.000    ?          0.000    ?          ?        0.000     0.200
                 0.000    0.000    ?          0.000    ?          ?        0.000     0.200
                 0.000    0.000    ?          0.000    ?          ?        0.000     0.200
                 0.000    0.000    ?          0.000    ?          ?        0.000     0.100
Weighted Avg.    0.300    0.300    ?          0.300    ?          ?        0.000     0.220

=== Confusion Matrix ===

 a b c d e   <-- classified as
 3 0 0 0 0 | a = a
 2 0 0 0 0 | b = b
 2 0 0 0 0 | c = c
 2 0 0 0 0 | d = d
 1 0 0 0 0 | e = e
```



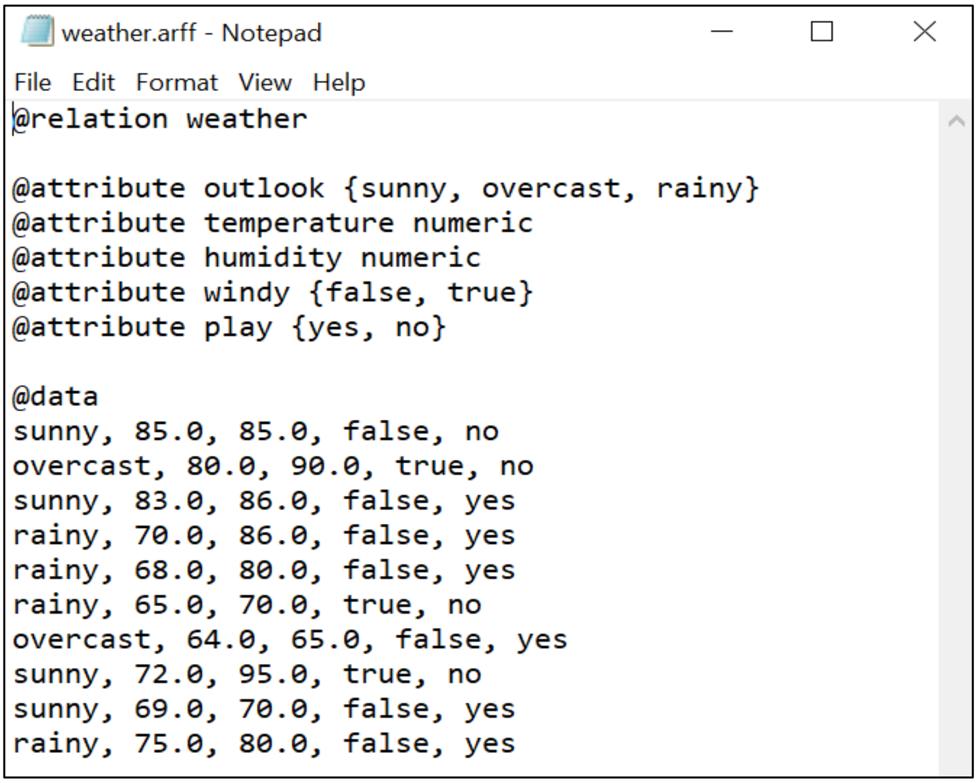
EM -I 100 -N -1 -X 10 -max -1 -ll-cv 1.0E-6 -ll-iter 1.0E-6 -M 1.0E-6 -K 10 -num-slots 1 -S 100

● Use training set
○ Supplied test set
○ Percentage split % 66
○ Classes to clusters evaluation
(Num) phone
☑ Store clusters for visualization

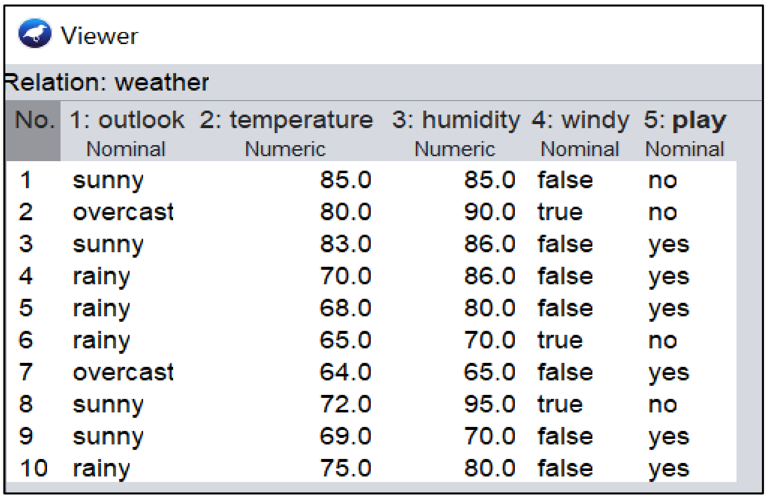
23:05:30 - EM

```
name
  a                    4
  b                    3
  c                    3
  d                    3
  e                    2
  [total]             15
id
  mean             105.5
  std. dev.       2.8723

salary
  low                  5
  medium               4
  high                 4
  [total]             13
exp
  mean               3.1
  std. dev.       1.4457

gender
  male                 7
  female               5
  [total]             12
phone
  mean      4890970727.3
  std. dev. 3246534202.8357
```

2) Create a Weather Table with training data set which includes attributes like outlook, temperature, humidity, windy, play with the help of Data Mining Tool WEKA.

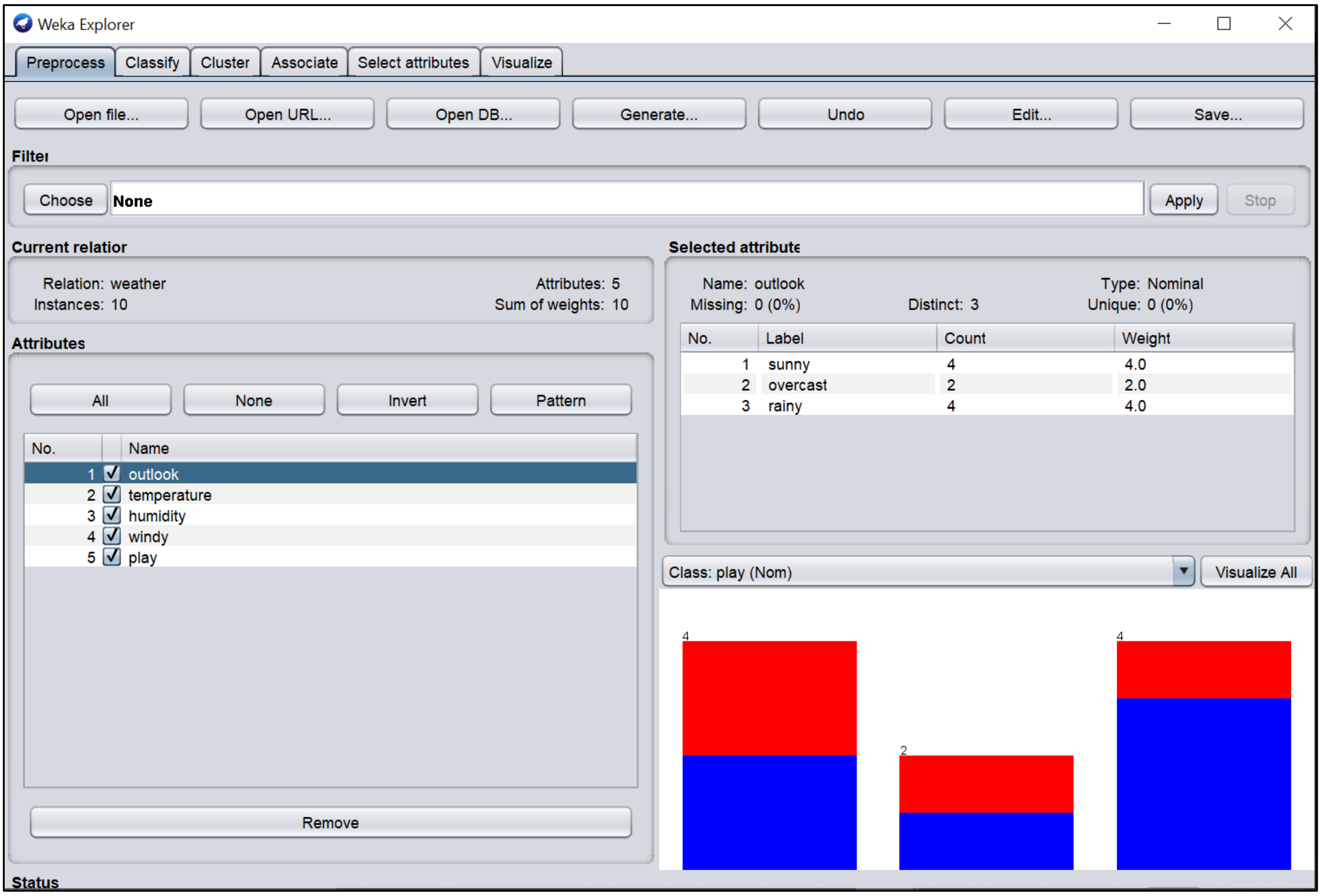


```
@relation weather

@attribute outlook {sunny, overcast, rainy}
@attribute temperature numeric
@attribute humidity numeric
@attribute windy {false, true}
@attribute play {yes, no}

@data
sunny, 85.0, 85.0, false, no
overcast, 80.0, 90.0, true, no
sunny, 83.0, 86.0, false, yes
rainy, 70.0, 86.0, false, yes
rainy, 68.0, 80.0, false, yes
rainy, 65.0, 70.0, true, no
overcast, 64.0, 65.0, false, yes
sunny, 72.0, 95.0, true, no
sunny, 69.0, 70.0, false, yes
rainy, 75.0, 80.0, false, yes
```



Relation: weather

No.	1: outlook	2: temperature	3: humidity	4: windy	5: play
	Nominal	Numeric	Numeric	Nominal	Nominal
1	sunny	85.0	85.0	false	no
2	overcast	80.0	90.0	true	no
3	sunny	83.0	86.0	false	yes
4	rainy	70.0	86.0	false	yes
5	rainy	68.0	80.0	false	yes
6	rainy	65.0	70.0	true	no
7	overcast	64.0	65.0	false	yes
8	sunny	72.0	95.0	true	no
9	sunny	69.0	70.0	false	yes
10	rainy	75.0	80.0	false	yes



Filter: None

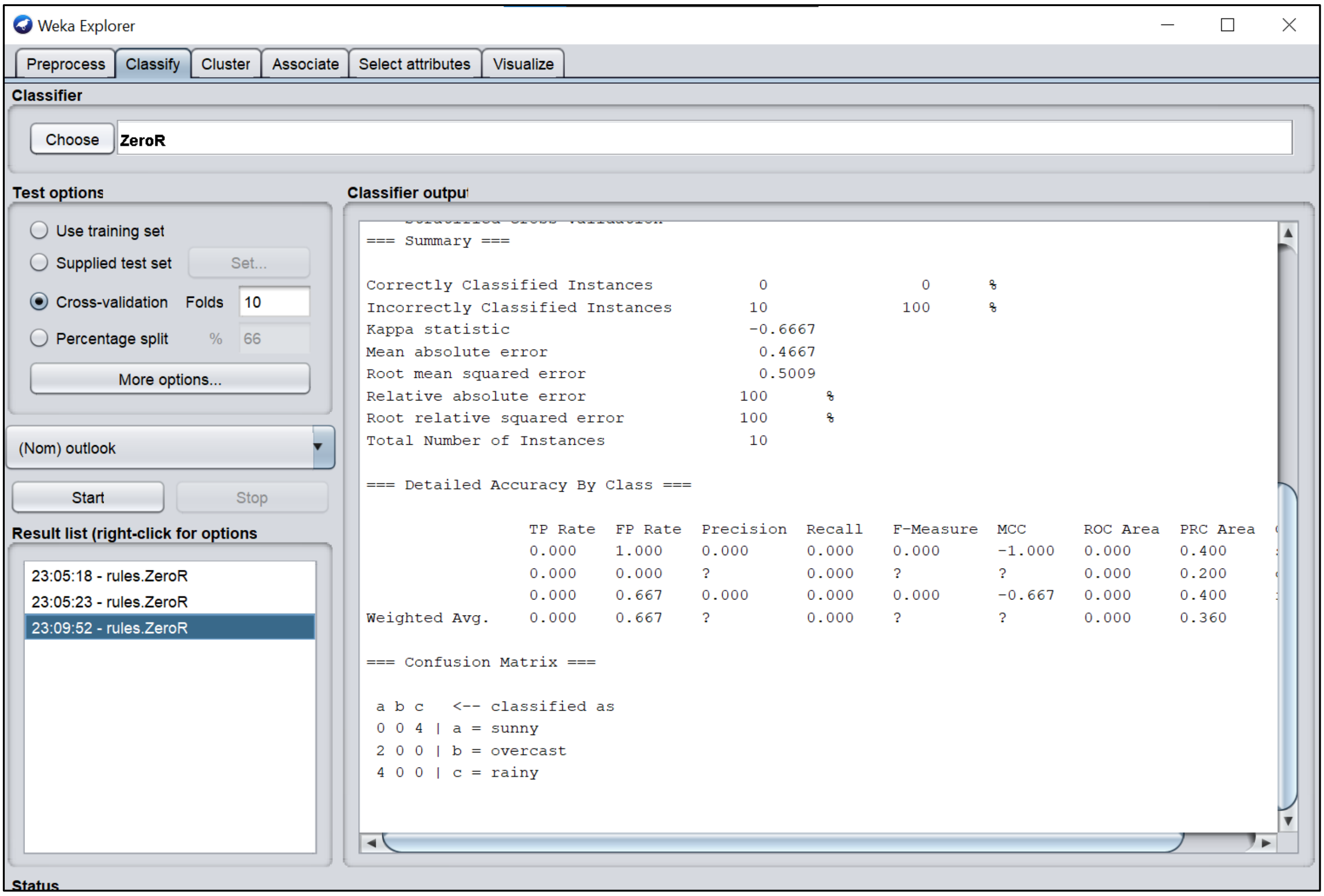
Relation: weather
Instances: 10
Attributes: 5
Sum of weights: 10

Name: outlook
Missing: 0 (0%)
Distinct: 3
Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	sunny	4	4.0
2	overcast	2	2.0
3	rainy	4	4.0

No.		Name
1	☑	outlook
2	☑	temperature
3	☑	humidity
4	☑	windy
5	☑	play

Class: play (Nom)



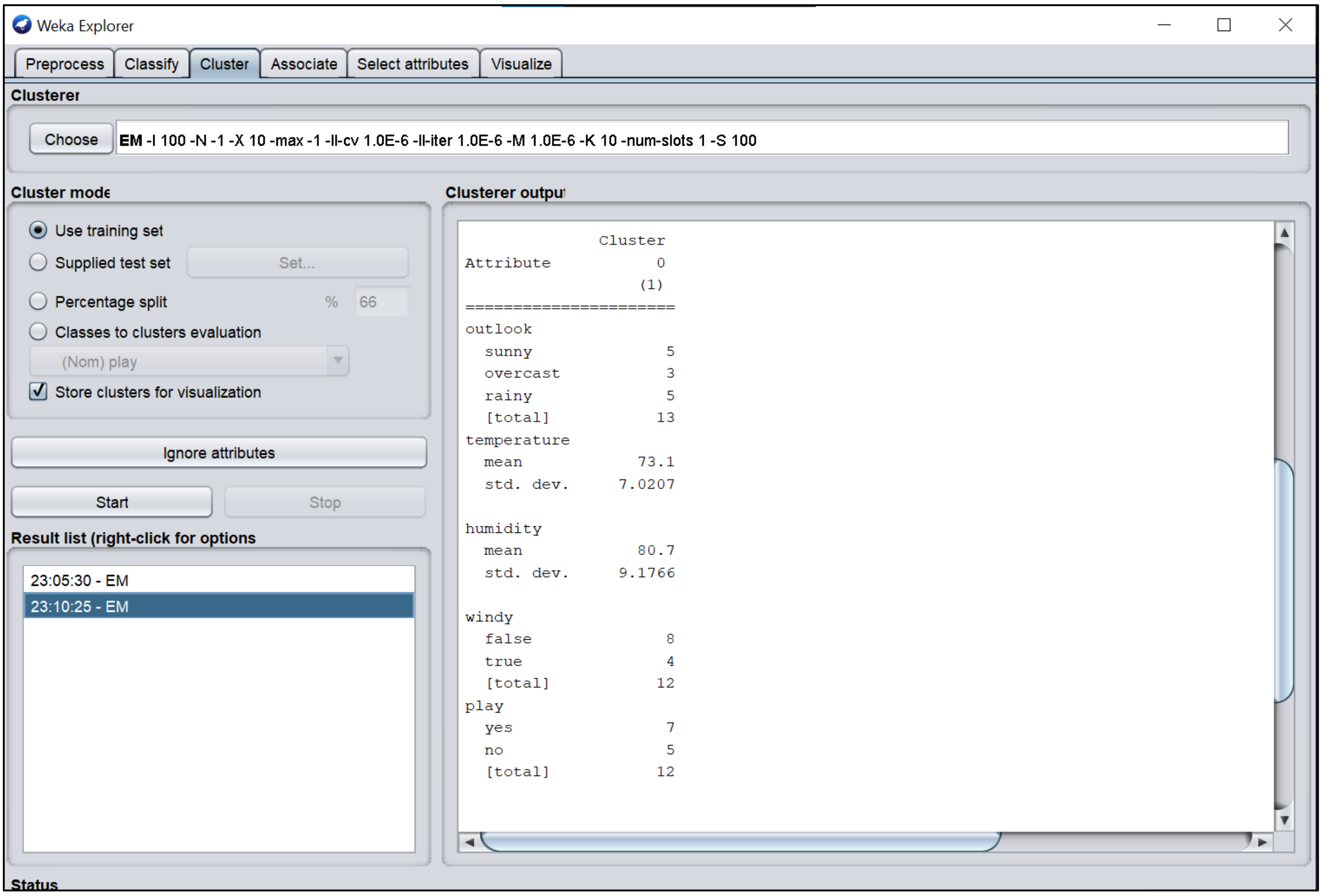
ZeroR

○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) outlook

23:05:18 - rules.ZeroR
23:05:23 - rules.ZeroR
23:09:52 - rules.ZeroR

```
=== Summary ===

Correctly Classified Instances           0                0      %
Incorrectly Classified Instances        10              100      %
Kappa statistic                         -0.6667
Mean absolute error                      0.4667
Root mean squared error                  0.5009
Relative absolute error                100      %
Root relative squared error            100      %
Total Number of Instances               10

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area
                 0.000    1.000    0.000      0.000    0.000      -1.000   0.000     0.400
                 0.000    0.000    ?          0.000    ?          ?        0.000     0.200
                 0.000    0.667    0.000      0.000    0.000      -0.667   0.000     0.400
Weighted Avg.    0.000    0.667    ?          0.000    ?          ?        0.000     0.360

=== Confusion Matrix ===

 a b c   <-- classified as
 0 0 4 | a = sunny
 2 0 0 | b = overcast
 4 0 0 | c = rainy
```



EM -I 100 -N -1 -X 10 -max -1 -ll-cv 1.0E-6 -ll-iter 1.0E-6 -M 1.0E-6 -K 10 -num-slots 1 -S 100

● Use training set
○ Supplied test set
○ Percentage split % 66
○ Classes to clusters evaluation
(Nom) play
☑ Store clusters for visualization

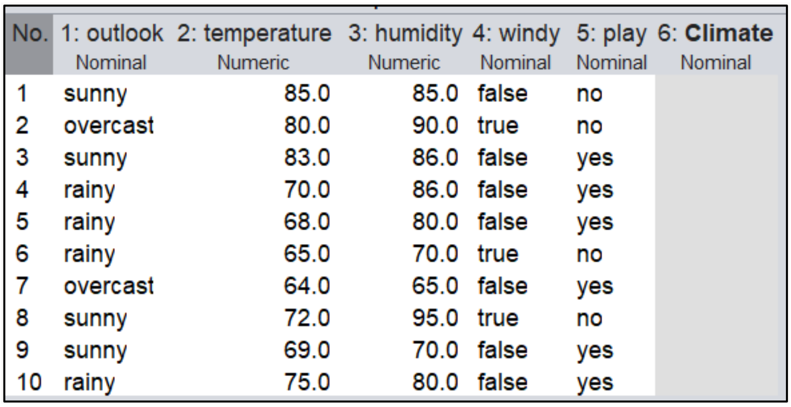
23:05:30 - EM
23:10:25 - EM

```
              Cluster
Attribute           0
                  (1)
======================
outlook
  sunny             5
  overcast          3
  rainy             5
  [total]          13
temperature
  mean           73.1
  std. dev.    7.0207

humidity
  mean           80.7
  std. dev.    9.1766

windy
  false             8
  true              4
  [total]          12
play
  yes               7
  no                5
  [total]          12
```

Q3) Apply Pre-Processing techniques to the training data set of Weather Table (Based on question 2.

i) Add

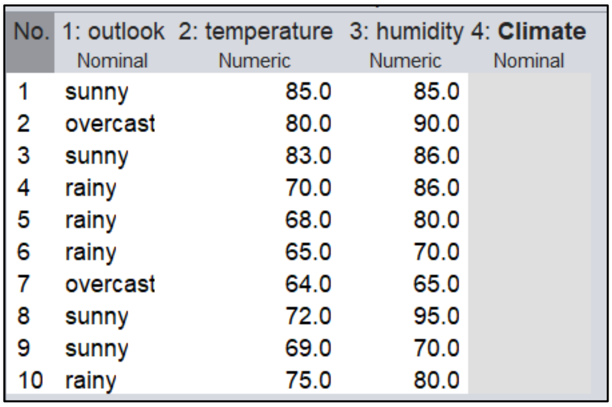
(Climate Attribute)



No.	1: outlook	2: temperature	3: humidity	4: windy	5: play	6: **Climate**
	Nominal	Numeric	Numeric	Nominal	Nominal	Nominal
1	sunny	85.0	85.0	false	no	
2	overcast	80.0	90.0	true	no	
3	sunny	83.0	86.0	false	yes	
4	rainy	70.0	86.0	false	yes	
5	rainy	68.0	80.0	false	yes	
6	rainy	65.0	70.0	true	no	
7	overcast	64.0	65.0	false	yes	
8	sunny	72.0	95.0	true	no	
9	sunny	69.0	70.0	false	yes	
10	rainy	75.0	80.0	false	yes	

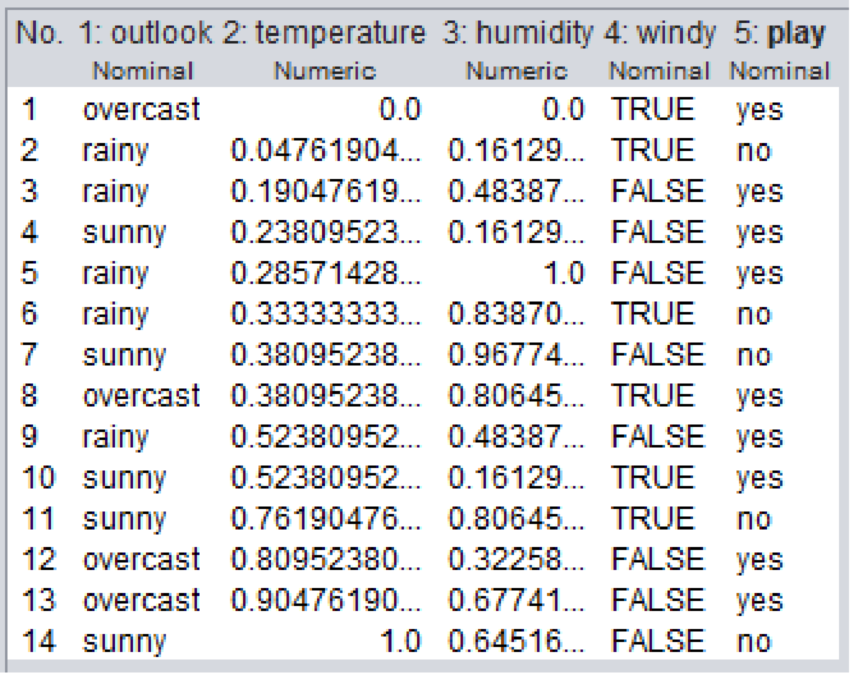
ii) Remove

(Windy and Play attribute)



No.	1: outlook	2: temperature	3: humidity	4: **Climate**
	Nominal	Numeric	Numeric	Nominal
1	sunny	85.0	85.0	
2	overcast	80.0	90.0	
3	sunny	83.0	86.0	
4	rainy	70.0	86.0	
5	rainy	68.0	80.0	
6	rainy	65.0	70.0	
7	overcast	64.0	65.0	
8	sunny	72.0	95.0	
9	sunny	69.0	70.0	
10	rainy	75.0	80.0	

iii) Normalization

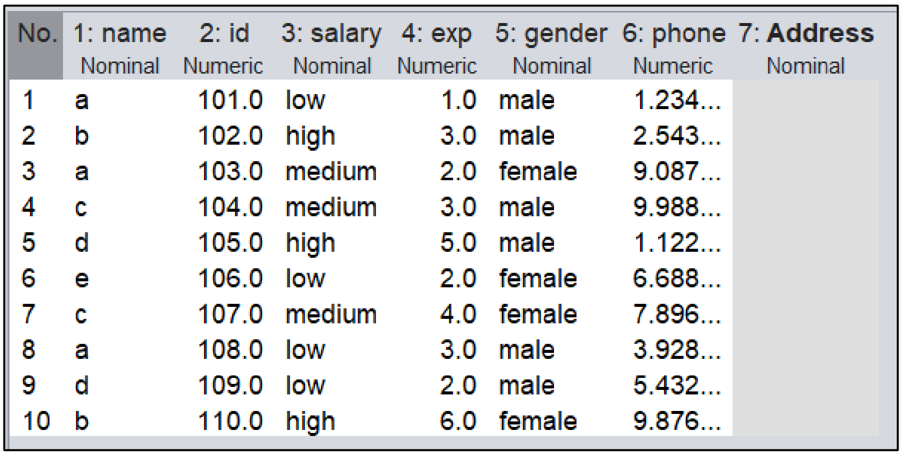


No.	1: outlook	2: temperature	3: humidity	4: windy	5: **play**
	Nominal	Numeric	Numeric	Nominal	Nominal
1	overcast	0.0	0.0	TRUE	yes
2	rainy	0.04761904...	0.16129...	TRUE	no
3	rainy	0.19047619...	0.48387...	FALSE	yes
4	sunny	0.23809523...	0.16129...	FALSE	yes
5	rainy	0.28571428...	1.0	FALSE	yes
6	rainy	0.33333333...	0.83870...	TRUE	no
7	sunny	0.38095238...	0.96774...	FALSE	no
8	overcast	0.38095238...	0.80645...	TRUE	yes
9	rainy	0.52380952...	0.48387...	FALSE	yes
10	sunny	0.52380952...	0.16129...	TRUE	yes
11	sunny	0.76190476...	0.80645...	TRUE	no
12	overcast	0.80952380...	0.32258...	FALSE	yes
13	overcast	0.90476190...	0.67741...	FALSE	yes
14	sunny	1.0	0.64516...	FALSE	no

Q4) Apply Pre-Processing techniques to the training data set of Employee Table (Based on question 1.

i) Add

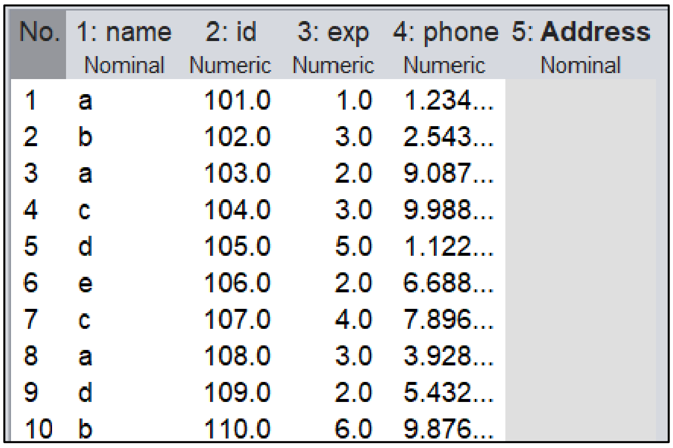
(Address Attribute)



No.	1: name	2: id	3: salary	4: exp	5: gender	6: phone	7: **Address**
	Nominal	Numeric	Nominal	Numeric	Nominal	Numeric	Nominal
1	a	101.0	low	1.0	male	1.234...	
2	b	102.0	high	3.0	male	2.543...	
3	a	103.0	medium	2.0	female	9.087...	
4	c	104.0	medium	3.0	male	9.988...	
5	d	105.0	high	5.0	male	1.122...	
6	e	106.0	low	2.0	female	6.688...	
7	c	107.0	medium	4.0	female	7.896...	
8	a	108.0	low	3.0	male	3.928...	
9	d	109.0	low	2.0	male	5.432...	
10	b	110.0	high	6.0	female	9.876...	

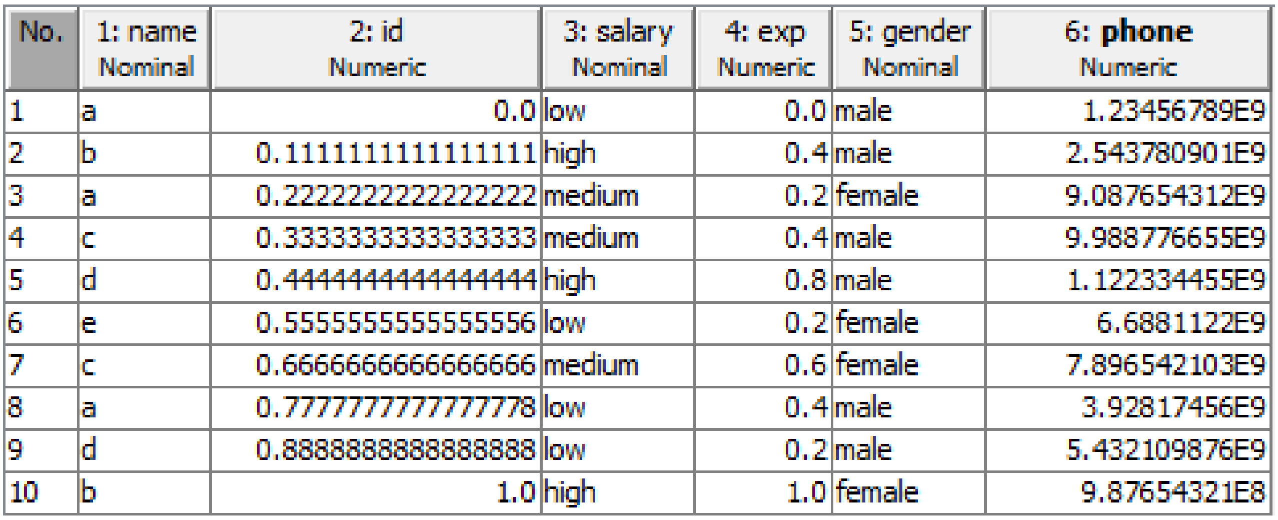
ii) Remove

(Salary and Gender Attribute)



No.	1: name	2: id	3: exp	4: phone	5: **Address**
	Nominal	Numeric	Numeric	Numeric	Nominal
1	a	101.0	1.0	1.234...	
2	b	102.0	3.0	2.543...	
3	a	103.0	2.0	9.087...	
4	c	104.0	3.0	9.988...	
5	d	105.0	5.0	1.122...	
6	e	106.0	2.0	6.688...	
7	c	107.0	4.0	7.896...	
8	a	108.0	3.0	3.928...	
9	d	109.0	2.0	5.432...	
10	b	110.0	6.0	9.876...	

iii) Normalization



No.	1: name	2: id	3: salary	4: exp	5: gender	6: **phone**
	Nominal	Numeric	Nominal	Numeric	Nominal	Numeric
1	a	0.0	low	0.0	male	1.23456789E9
2	b	0.1111111111111111	high	0.4	male	2.543780901E9
3	a	0.2222222222222222	medium	0.2	female	9.087654312E9
4	c	0.3333333333333333	medium	0.4	male	9.988776655E9
5	d	0.4444444444444444	high	0.8	male	1.122334455E9
6	e	0.5555555555555556	low	0.2	female	6.6881122E9
7	c	0.6666666666666666	medium	0.6	female	7.896542103E9
8	a	0.7777777777777778	low	0.4	male	3.92817456E9
9	d	0.8888888888888888	low	0.2	male	5.432109876E9
10	b	1.0	high	1.0	female	9.87654321E8