Homework 7

Clustering with WEKA

Ομάδα G

Ευαγγέλου Σωτήρης 2159

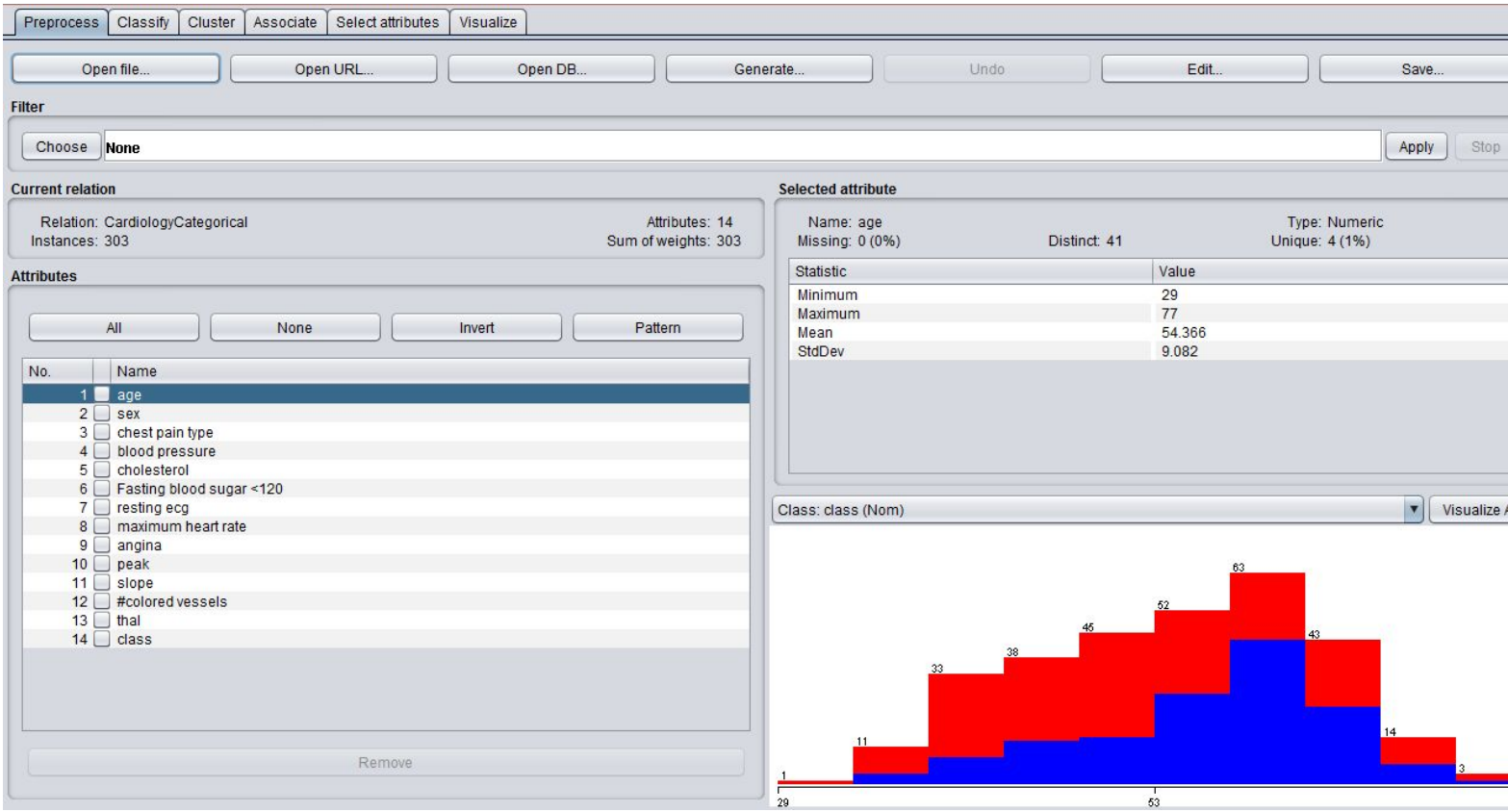
Καλαής Κωσταντίνος 2146

Χατζηευφραιμίδης Λευτέρης 2209

UNIVERSITY OF THESSALY

(2.1)

- We load the "*CardiologyCategorical.csv*" data file to build a neural network (using the MLP algorithm) that predicts whether a patient has a heart condition.



- We choose "*True*" value on "*normalizeAttributes*" to normalize our features and set learning rate for weights updates to 0.01

- The 10-fold cross-validation accuracy of your model (A1) is 84.1584%

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

Classifier

Choose **MultilayerPerceptron** -L 0.01 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a

Test options

- Use training set
- Supplied test set — Set...
- Cross-validation Folds 10
- Percentage split % 70

More options...

(Nom) class

Start | Stop

Result list (right-click for options)

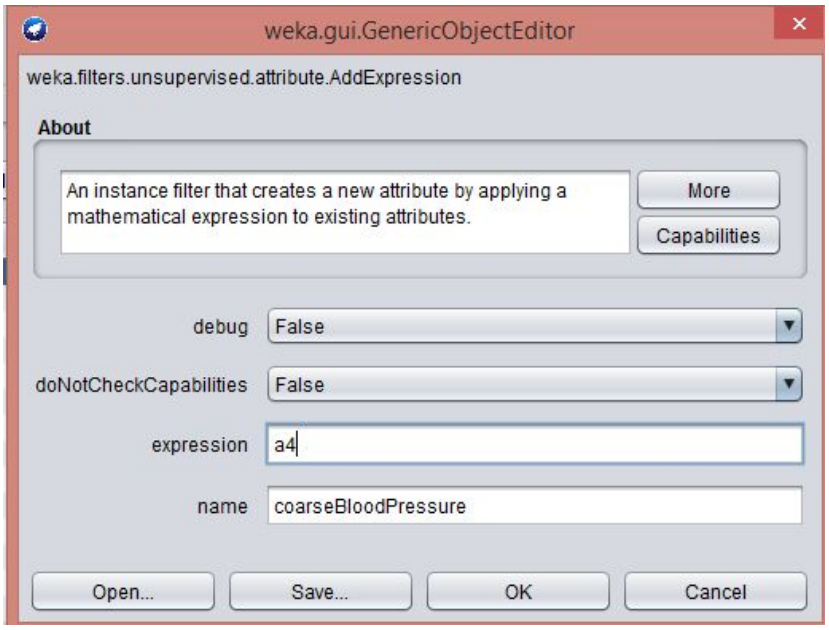
14:56:01 - functions.MultilayerPerceptron

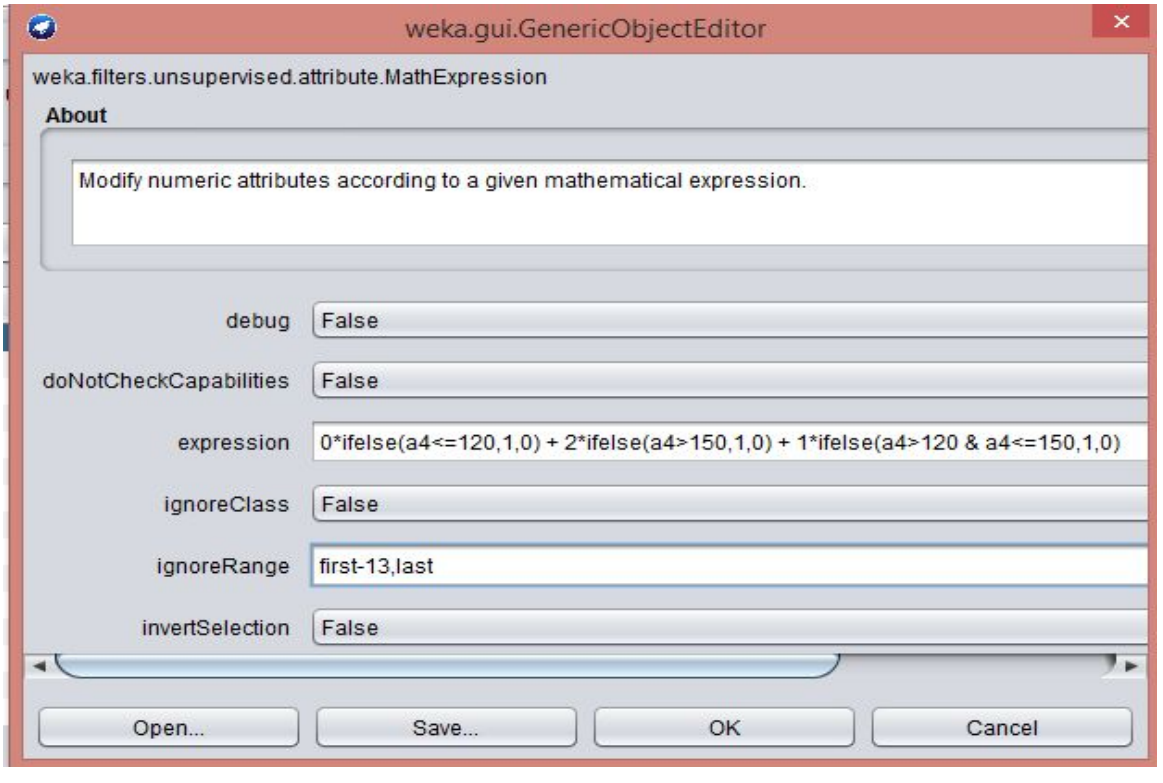
Classifier output

```
    Node 1


Time taken to build model: 1.3 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         255               84.1584 %
Incorrectly Classified Instances        48               15.8416 %
Kappa statistic                          0.6795
Mean absolute error                      0.2013
Root mean squared error                  0.345
Relative absolute error                 40.5837 %
Root relative squared error             69.2689 %
Total Number of Instances              303

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,804    0,127    0,841      0,804    0,822      0,680    0,909     0,901     Sick
                 0,873    0,196    0,842      0,873    0,857      0,680    0,909     0,908     Healthy
Weighted Avg.    0,842    0,165    0,842      0,842    0,841      0,680    0,909     0,905

=== Confusion Matrix ===

   a   b   <-- classified as
 111  27 |   a = Sick
  21 144 |   b = Healthy
```

(2.2)

- We create a new attribute *coarseBloodPressure*, with values: Low if *blood pressure* is less than or equal to 120, Normal if *blood pressure* is greater than 120 but less than or equal to 150, and High if *blood pressure* is greater than 150
- First we create a copy of *blood pressure* with the new name using filter *AddExpression* and then with *MathExpression* we transform *blood pressure's* values as follows :



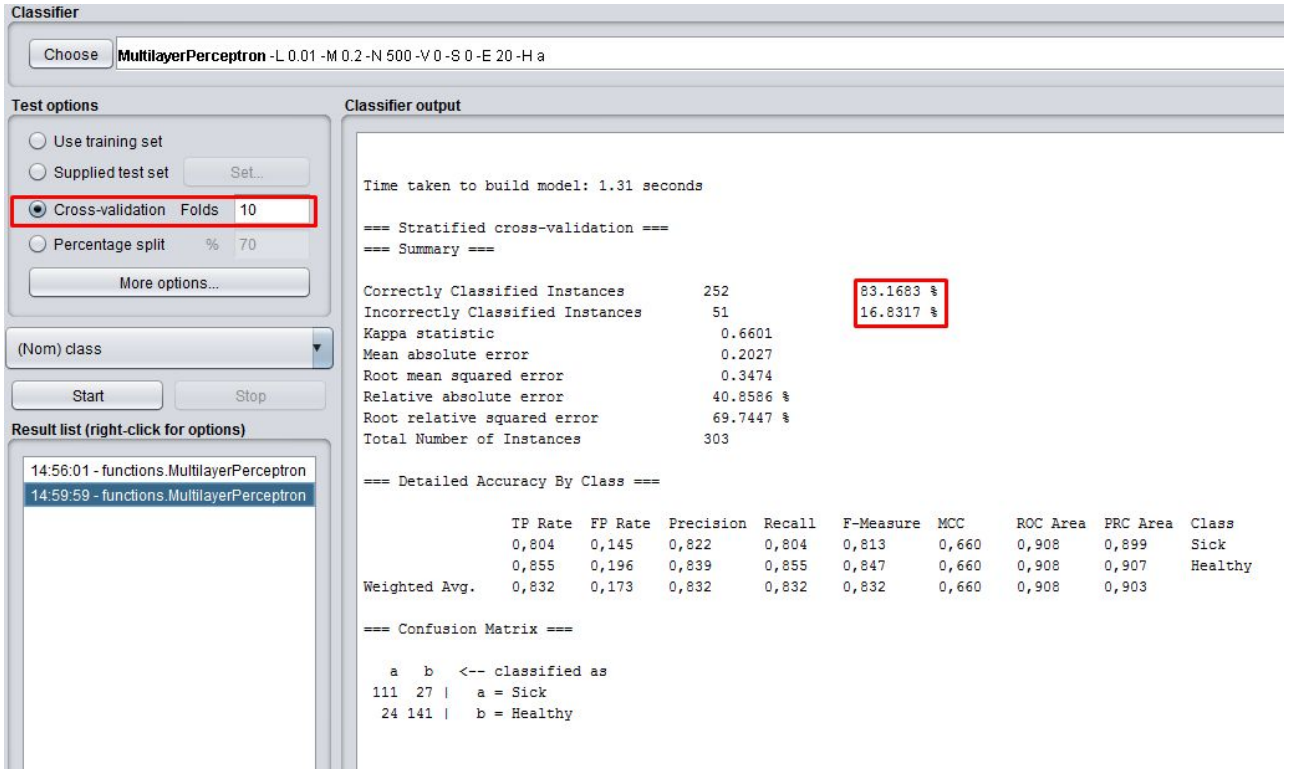


1: age	2: sex	3: chest pain type	4: blood pressure	5: cholesterol	6: Fasting blood sugar 120	7: resting ecg	8: maximum heart rate	9: angina	10: peak	11: slope	12: #colored vessels	13: thal	14: coarseBloodPressure	15: class
Numeric	Nominal	Nominal	Numeric	Numeric	Nominal	Nominal	Numeric	Nominal	Numeric	Nominal	Numeric	Nominal	Numeric	Nominal
60.0	Male	Asymptomatic	130.0	206.0	FALSE	Hyp	132.0	TRUE	2.4	Flat	2.0	Rev	(Normal)	1.0 Sick
49.0	Male	Abnormal Angi...	130.0	266.0	FALSE	Normal	171.0	FALSE	0.6	Up	0.0	Nor...		1.0 Healthy
64.0	Male	Angina	110.0	211.0	FALSE	Hyp	144.0	TRUE	1.8	Flat	0.0	Nor...	(Low)	0.0 Healthy
63.0	Male	Asymptomatic	130.0	254.0	FALSE	Hyp	147.0	FALSE	1.4	Flat	1.0	Rev		1.0 Sick
53.0	Male	Asymptomatic	140.0	203.0	TRUE	Hyp	155.0	TRUE	3.1	Down	0.0	Rev		1.0 Sick
58.0	Fem...	Angina	150.0	283.0	TRUE	Hyp	162.0	FALSE	1.0	Up	0.0	Nor...		1.0 Healthy
58.0	Male	Abnormal Angi...	120.0	284.0	FALSE	Hyp	160.0	FALSE	1.8	Flat	0.0	Nor...		0.0 Sick
58.0	Male	NoTang	132.0	224.0	FALSE	Hyp	173.0	FALSE	3.2	Up	2.0	Rev		1.0 Sick
63.0	Male	Angina	145.0	233.0	TRUE	Hyp	150.0	FALSE	2.3	Down	0.0	Fix		1.0 Healthy
67.0	Male	Asymptomatic	160.0	286.0	FALSE	Hyp	108.0	TRUE	1.5	Flat	3.0	Nor...	(High)	2.0 Sick
67.0	Male	Asymptomatic	120.0	229.0	FALSE	Hyp	129.0	TRUE	2.6	Flat	2.0	Rev		0.0 Sick
37.0	Male	NoTang	130.0	250.0	FALSE	Normal	187.0	FALSE	3.5	Down	0.0	Nor...		1.0 Healthy
41.0	Fem...	Abnormal Angi...	130.0	204.0	FALSE	Hyp	172.0	FALSE	1.4	Up	0.0	Nor...		1.0 Healthy
56.0	Male	Abnormal Angi...	120.0	236.0	FALSE	Normal	178.0	FALSE	0.8	Up	0.0	Nor...		0.0 Healthy
62.0	Fem...	Asymptomatic	140.0	268.0	FALSE	Hyp	160.0	FALSE	3.6	Down	2.0	Nor...		1.0 Sick
57.0	Fem...	Asymptomatic	120.0	354.0	FALSE	Normal	163.0	TRUE	0.6	Up	0.0	Nor...		0.0 Healthy
57.0	Male	Asymptomatic	140.0	192.0	FALSE	Normal	148.0	FALSE	0.4	Flat	0.0	Fix		1.0 Healthy
56.0	Fem...	Abnormal Angi...	140.0	294.0	FALSE	Hyp	153.0	FALSE	1.3	Flat	0.0	Nor...		1.0 Healthy
56.0	Male	NoTang	130.0	256.0	TRUE	Hyp	142.0	TRUE	0.6	Flat	1.0	Fix		1.0 Sick
44.0	Male	Abnormal Angi...	120.0	263.0	FALSE	Normal	173.0	FALSE	0.0	Up	0.0	Rev		0.0 Healthy
50.0	Fem...	NoTang	120.0	219.0	FALSE	Normal	158.0	FALSE	1.6	Flat	0.0	Nor...		0.0 Healthy
58.0	Fem...	NoTang	120.0	340.0	FALSE	Normal	172.0	FALSE	0.0	Up	0.0	Nor...		0.0 Healthy
66.0	Fem...	Angina	150.0	226.0	FALSE	Normal	114.0	FALSE	2.6	Down	0.0	Nor...		1.0 Healthy
43.0	Male	Asymptomatic	150.0	247.0	FALSE	Normal	171.0	FALSE	1.5	Up	0.0	Nor...		1.0 Healthy
40.0	Male	Asymptomatic	110.0	167.0	FALSE	Hyp	114.0	TRUE	2.0	Flat	0.0	Rev		0.0 Sick
69.0	Fem...	Angina	140.0	239.0	FALSE	Normal	151.0	FALSE	1.8	Up	2.0	Nor...		1.0 Healthy
60.0	Male	Asymptomatic	117.0	230.0	TRUE	Normal	160.0	TRUE	1.4	Up	2.0	Rev		0.0 Sick
64.0	Male	NoTang	140.0	335.0	FALSE	Normal	158.0	FALSE	0.0	Up	0.0	Nor...		1.0 Sick
59.0	Male	Asymptomatic	135.0	234.0	FALSE	Normal	161.0	FALSE	0.5	Flat	0.0	Rev		1.0 Healthy
44.0	Male	NoTang	130.0	233.0	FALSE	Normal	179.0	TRUE	0.4	Up	0.0	Nor...		1.0 Healthy
42.0	Male	Asymptomatic	140.0	226.0	FALSE	Normal	178.0	FALSE	0.0	Up	0.0	Nor...		1.0 Healthy
43.0	Male	Asymptomatic	120.0	177.0	FALSE	Hyp	120.0	TRUE	2.5	Flat	0.0	Rev		0.0 Sick

(2.3)

- We build a neural network (using the MLP algorithm) that predicts whether a patient has a heart condition, using the new attribute *coarseBloodPressure* instead of the original *blood pressure*. The 10-fold cross-validation accuracy of your model (A2) is 83.1683%

No.		Name
1	☐	age
2	☐	sex
3	☐	chest pain type
4	☐	cholesterol
5	☐	Fasting blood sugar <120
6	☐	resting ecg
7	☐	maximum heart rate
8	☐	angina
9	☐	peak
10	☐	slope
11	☐	#colored vessels
12	☐	thal
13	☐	coarseBloodPressure
14	☐	class

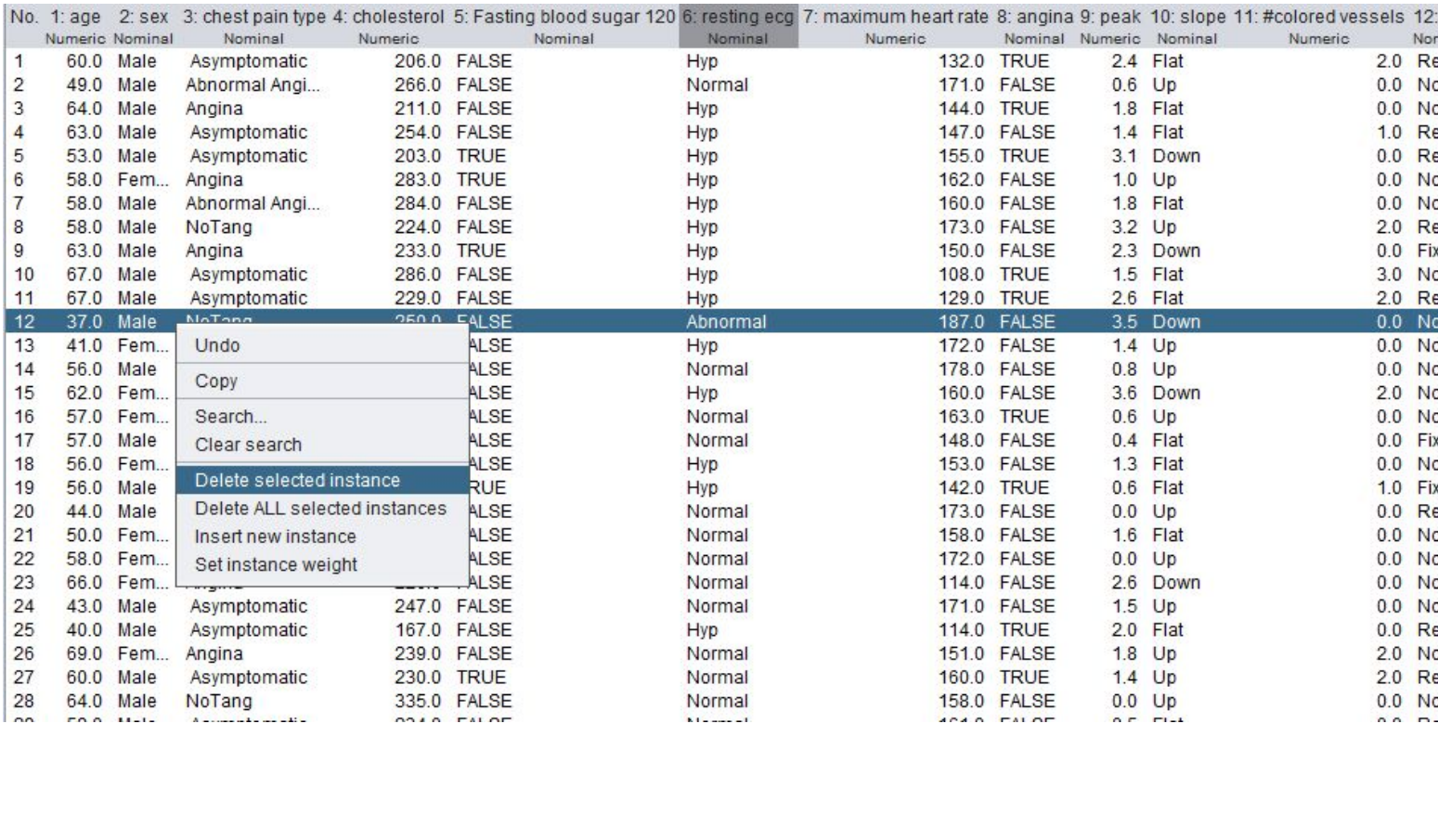


(2.4)

- The accuracies we achieve from using blood pressure in the first occasion and coarse blood pressure in the second do not have a significant difference (1%). The generalisation of the blood pressure feature does not seem to work better than raw values, but it usually depends to the problem.

(2.5)

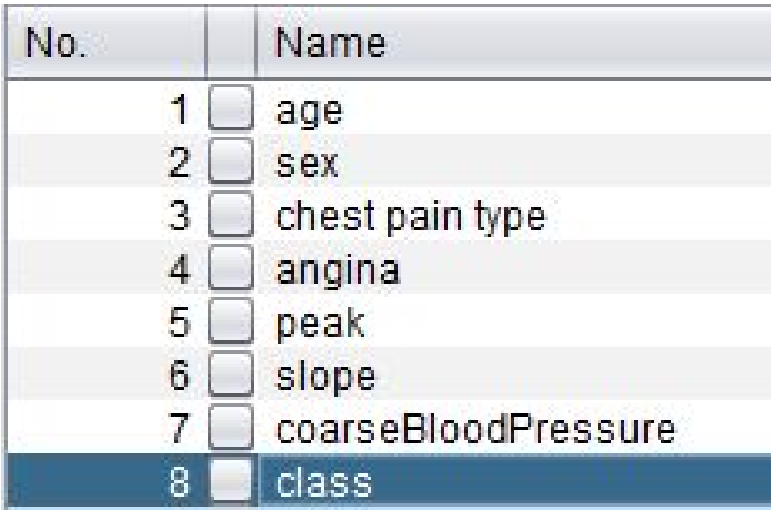
- We remove all records (4 in total) whose value of the attribute *resting ecg* is Abnormal.



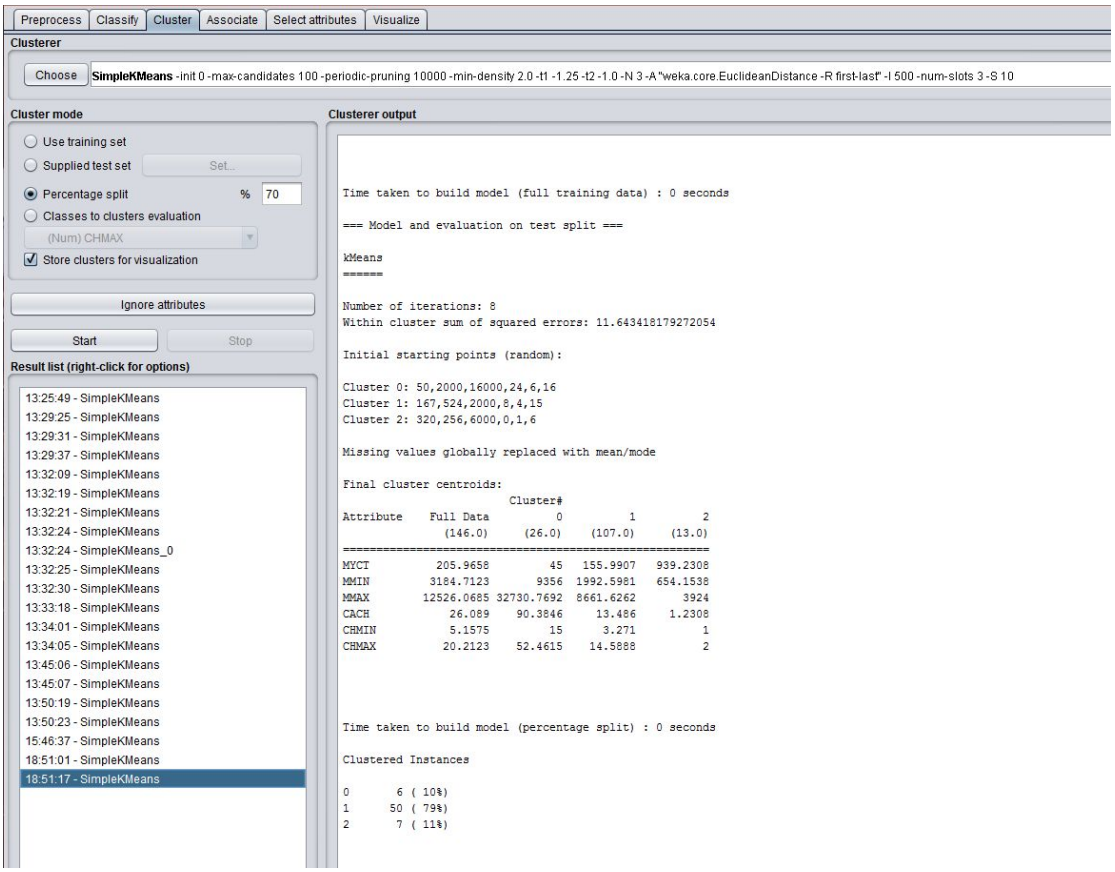
No.	1: age	2: sex	3: chest pain type	4: cholesterol	5: Fasting blood sugar 120	6: resting ecg	7: maximum heart rate	8: angina	9: peak	10: slope	11: #colored vessels	12:
	Numeric	Nominal	Nominal	Numeric	Nominal	Nominal	Numeric	Nominal	Numeric	Nominal	Numeric	Nor
1	60.0	Male	Asymptomatic	206.0	FALSE	Hyp	132.0	TRUE	2.4	Flat	2.0	Re
2	49.0	Male	Abnormal Angi...	266.0	FALSE	Normal	171.0	FALSE	0.6	Up	0.0	No
3	64.0	Male	Angina	211.0	FALSE	Hyp	144.0	TRUE	1.8	Flat	0.0	No
4	63.0	Male	Asymptomatic	254.0	FALSE	Hyp	147.0	FALSE	1.4	Flat	1.0	Re
5	53.0	Male	Asymptomatic	203.0	TRUE	Hyp	155.0	TRUE	3.1	Down	0.0	Re
6	58.0	Fem...	Angina	283.0	TRUE	Hyp	162.0	FALSE	1.0	Up	0.0	No
7	58.0	Male	Abnormal Angi...	284.0	FALSE	Hyp	160.0	FALSE	1.8	Flat	0.0	No
8	58.0	Male	NoTang	224.0	FALSE	Hyp	173.0	FALSE	3.2	Up	2.0	Re
9	63.0	Male	Angina	233.0	TRUE	Hyp	150.0	FALSE	2.3	Down	0.0	Fix
10	67.0	Male	Asymptomatic	286.0	FALSE	Hyp	108.0	TRUE	1.5	Flat	3.0	No
11	67.0	Male	Asymptomatic	229.0	FALSE	Hyp	129.0	TRUE	2.6	Flat	2.0	Re
12	37.0	Male	NoTang	250.0	FALSE	Abnormal	187.0	FALSE	3.5	Down	0.0	No
13	41.0	Fem...			ALSE	Hyp	172.0	FALSE	1.4	Up	0.0	No
14	56.0	Male			ALSE	Normal	178.0	FALSE	0.8	Up	0.0	No
15	62.0	Fem...			ALSE	Hyp	160.0	FALSE	3.6	Down	2.0	No
16	57.0	Fem...			ALSE	Normal	163.0	TRUE	0.6	Up	0.0	No
17	57.0	Male			ALSE	Normal	148.0	FALSE	0.4	Flat	0.0	Fix
18	56.0	Fem...			ALSE	Hyp	153.0	FALSE	1.3	Flat	0.0	No
19	56.0	Male			RUE	Hyp	142.0	TRUE	0.6	Flat	1.0	Fix
20	44.0	Male			ALSE	Normal	173.0	FALSE	0.0	Up	0.0	Re
21	50.0	Fem...			ALSE	Normal	158.0	FALSE	1.6	Flat	0.0	No
22	58.0	Fem...			ALSE	Normal	172.0	FALSE	0.0	Up	0.0	No
23	66.0	Fem...			ALSE	Normal	114.0	FALSE	2.6	Down	0.0	No
24	43.0	Male	Asymptomatic	247.0	FALSE	Normal	171.0	FALSE	1.5	Up	0.0	No
25	40.0	Male	Asymptomatic	167.0	FALSE	Hyp	114.0	TRUE	2.0	Flat	0.0	Re
26	69.0	Fem...	Angina	239.0	FALSE	Normal	151.0	FALSE	1.8	Up	2.0	No
27	60.0	Male	Asymptomatic	230.0	TRUE	Normal	160.0	TRUE	1.4	Up	2.0	Re
28	64.0	Male	NoTang	335.0	FALSE	Normal	158.0	FALSE	0.0	Up	0.0	No

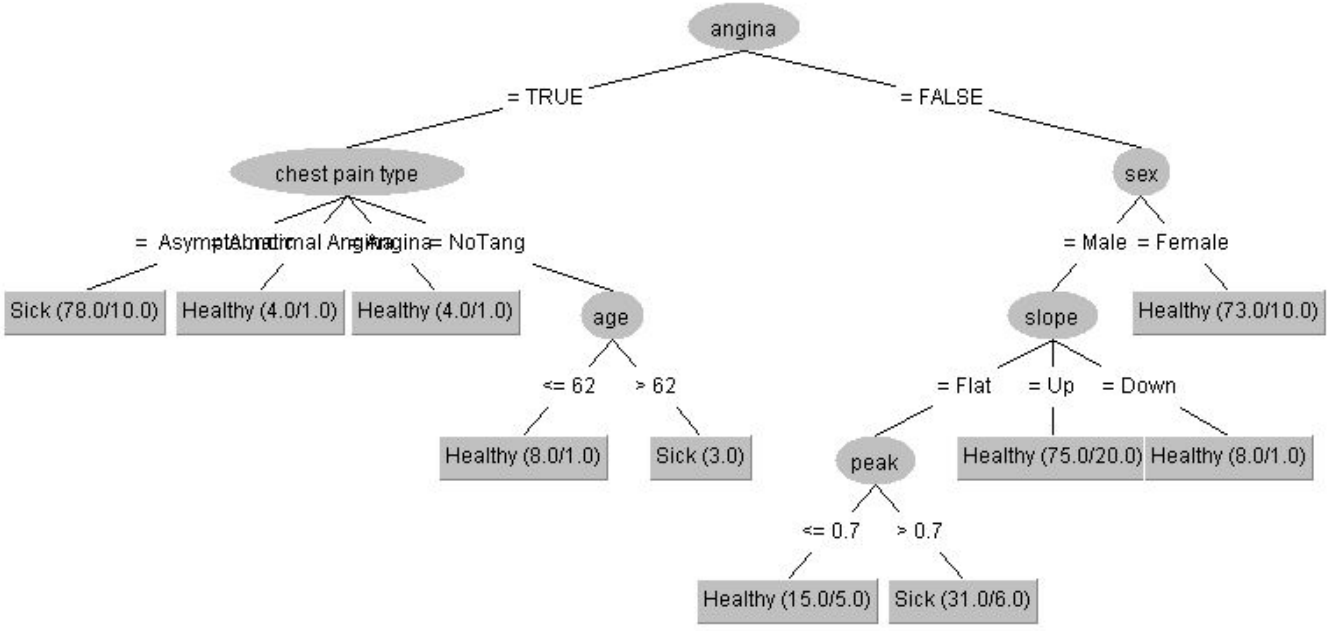
(2.6)

- We construct a decision tree (using the J48 Algorithm) that predicts whether a patient has a heart condition, given the attributes *age*, *sex*, *chest pain type*, *coarseBloodPressure*, *angina*, *peak*, and *slope*.



No.		Name
1	☐	age
2	☐	sex
3	☐	chest pain type
4	☐	angina
5	☐	peak
6	☐	slope
7	☐	coarseBloodPressure
8	☐	class


Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Clusterer
Choose
SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 3 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 3 -S 10
Cluster mode
Use training set
Supplied test set
Set...
Percentage split
%
70
Classes to clusters evaluation
(Num) CHMAX
Store clusters for visualization
Ignore attributes
Start
Stop
Result list (right-click for options)
13:25:49 - SimpleKMeans
13:29:25 - SimpleKMeans
13:29:31 - SimpleKMeans
13:29:37 - SimpleKMeans
13:32:09 - SimpleKMeans
13:32:19 - SimpleKMeans
13:32:21 - SimpleKMeans
13:32:24 - SimpleKMeans
13:32:24 - SimpleKMeans_0
13:32:25 - SimpleKMeans
13:32:30 - SimpleKMeans
13:33:18 - SimpleKMeans
13:34:01 - SimpleKMeans
13:34:05 - SimpleKMeans
13:45:06 - SimpleKMeans
13:45:07 - SimpleKMeans
13:50:19 - SimpleKMeans
13:50:23 - SimpleKMeans
15:46:37 - SimpleKMeans
18:51:01 - SimpleKMeans
18:51:17 - SimpleKMeans
Clusterer output
Time taken to build model (full training data) : 0 seconds
=== Model and evaluation on test split ===
kMeans
======
Number of iterations: 8
Within cluster sum of squared errors: 11.643418179272054
Initial starting points (random):
Cluster 0: 50,2000,16000,24,6,16
Cluster 1: 167,524,2000,8,4,15
Cluster 2: 320,256,6000,0,1,6
Missing values globally replaced with mean/mode
Final cluster centroids:
Cluster#
Attribute Full Data 0 1 2
(146.0) (26.0) (107.0) (13.0)
MYCT 205.9658 45 155.9907 939.2308
MMIN 3184.7123 9356 1992.5981 654.1538
MMAX 12526.0685 32730.7692 8661.6262 3924
CACH 26.089 90.3846 13.486 1.2308
CHMIN 5.1575 15 3.271 1
CHMAX 20.2123 52.4615 14.5888 2
Time taken to build model (percentage split) : 0 seconds
Clustered Instances
0 6 (10%)
1 50 (79%)
2 7 (11%)


angina
= TRUE
= FALSE
chest pain type
sex
= Male
= Female
Healthy (73.0/10.0)
slope
= Flat
= Up
= Down
Healthy (75.0/20.0)
Healthy (8.0/1.0)
peak
<= 0.7
> 0.7
Healthy (15.0/5.0)
Sick (31.0/6.0)
Sick (78.0/10.0)
Healthy (4.0/1.0)
Healthy (4.0/1.0)
age
<= 62
> 62
Healthy (8.0/1.0)
Sick (3.0)

- Confusion matrix obtained with 10-fold cross-validation in our report.

=== Confusion Matrix ===

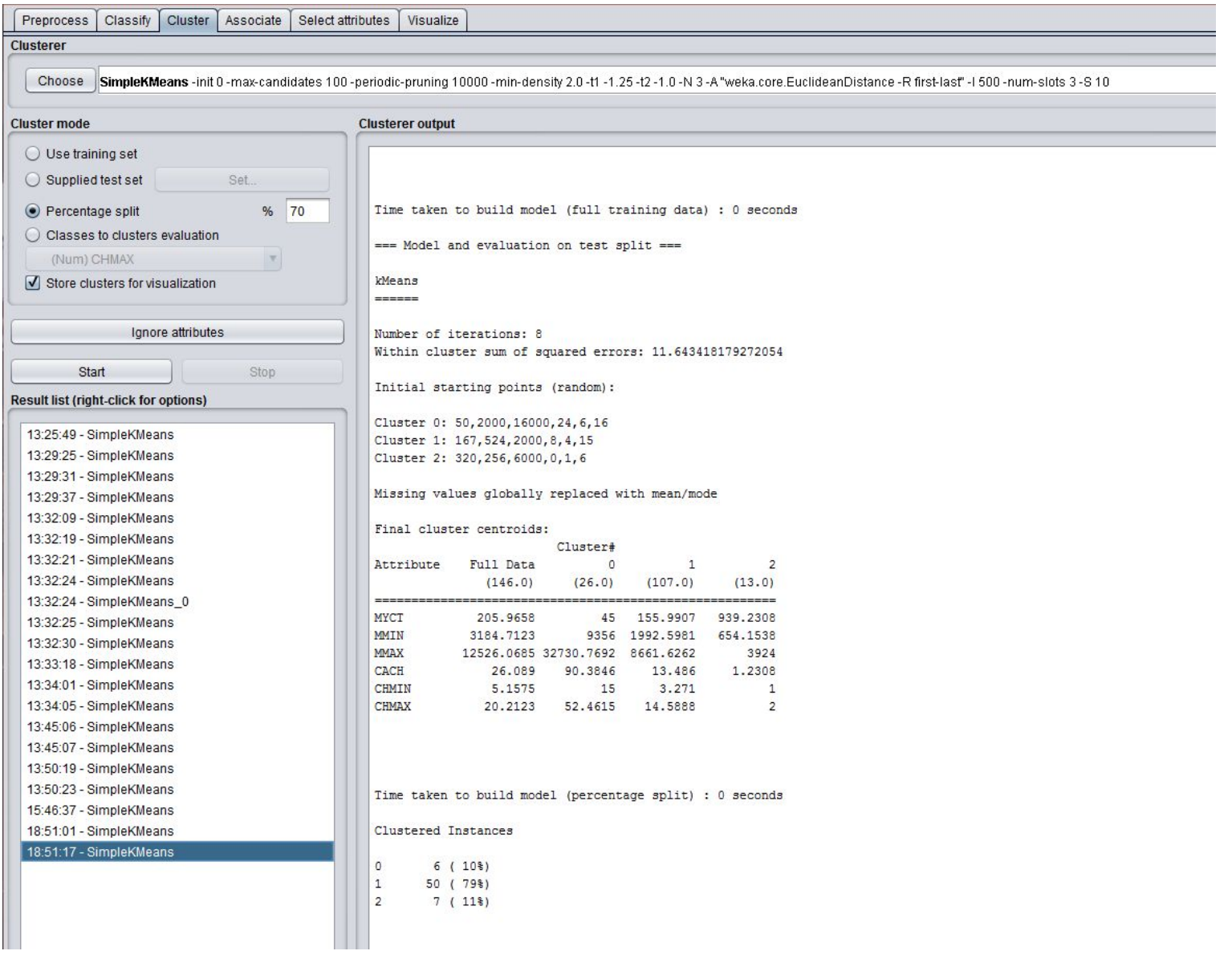
a b <-- classified as

23 8 | a = Sick

9 20 | b = Healthy

(3.1)

- We load the "*cpu.arff* " data file, cluster the data (using the simple k-Means algorithm, with k=3) and delete the attribute *class.*



- We can observe, that the main mass of the data instances belongs to one cluster (79%). The other two clusters include

a very small percentage of the examples that we can categorise as outliers.

(3.2)

- We discretize the attributes *MMIN*, *MMAX*, *CACH*, *CHMIN* and *CHMAX* using 3 buckets in one step and converting numeric values into nominal.

- Then, we apply the ‘NumericToNominal’ filter to every attribute in order to turn them to nominals.
- Then we find associations among these attributes using the Apriori algorithm, with support 0.2, confidence 0.95 and top 3 rules only being displayed. The result buffer :

```
=== Run information ===

Scheme:       weka.associations.Apriori -N 3 -T 0 -C 0.95 -D 0.05 -U 1.0 -M 0.2 -S -1.0 -c -1
Relation:
        cpu-weka.filters.unsupervised.attribute.Remove-R7-weka.filters.unsupervised.att
        ribute.Remove-R1-weka.filters.unsupervised.attribute.Discretize-B3-M-1.0-Rfirst-
        last-precision1-weka.filters.unsupervised.attribute.NumericToNominal-Rfirst-last
Instances:    209
Attributes:   5
         MMIN
         MMAX
         CACH
         CHMIN
         CHMAX
=== Associator model (full training set) ===


Apriori
=======

Minimum support: 0.9 (188 instances)
Minimum metric <confidence>: 0.95
Number of cycles performed: 2

Generated sets of large itemsets:

Size of set of large itemsets L(1): 3

Size of set of large itemsets L(2): 3

Size of set of large itemsets L(3): 1

Best rules found:
```

1. CACH='(-inf-85.3]' 194 ==> CHMIN='(-inf-17.3]' 193 <conf:(0.99)> lift:(1.02) lev:(0.02) [3] conv:(2.32)
2. MMIN='(-inf-10709.3]' CACH='(-inf-85.3]' 189 ==> CHMIN='(-inf-17.3]' 188 <conf:(0.99)> lift:(1.02) lev:(0.02) [3] conv:(2.26)
3. MMIN='(-inf-10709.3]' 201 ==> CHMIN='(-inf-17.3]' 197 <conf:(0.98)> lift:(1) lev:(0) [0] conv:(0.96)

(3.3)

- Using the original data, we list the eigenvalues associated with the attributes selected by the Principal Components Analysis method, when the amount of variance covered by the subset of attributes is 75%.

=== Run information ===

Evaluator: weka.attributeSelection.PrincipalComponents -R 0.75 -A 5
Search: weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1
Relation: cpu
Instances: 209
Attributes: 7
MYCT
MMIN
MMAX
CACH
CHMIN
CHMAX
class
Evaluation mode: evaluate on all training data

=== Attribute Selection on all input data ===

Search Method:
Attribute ranking.

Attribute Evaluator (unsupervised):

Principal Components Attribute Transformer

Correlation matrix

1	-0.34	-0.38	-0.32	-0.3	-0.25
-0.34	1	0.76	0.53	0.52	0.27
-0.38	0.76	1	0.54	0.56	0.53
-0.32	0.53	0.54	1	0.58	0.49
-0.3	0.52	0.56	0.58	1	0.55
-0.25	0.27	0.53	0.49	0.55	1

eigenvalue	proportion	cumulative
3.35674	0.55946	0.55946
-0.469MMAX-0.435CHMIN-0.429CACH-0.427MMIN-0.374CHMAX...		
0.82936	0.13823	0.69768
0.682MYCT+0.559CHMAX-0.333MMIN+0.275CHMIN+0.152CACH...		
0.73923	0.1232	0.82089
0.669MYCT+0.548MMIN-0.426CHMAX+0.264MMAX-0.03CHMIN...		

Eigenvectors

V1	V2	V3		
0.29	0.6822	0.6686	MYCT	
-0.4274		-0.333	0.5477	MMIN
-0.4691		-0.1141	0.2643	MMAX
-0.4286		0.1516	0.0199	CACH
-0.4353		0.2746	-0.0302	CHMIN
-0.3742		0.5588	-0.4264	CHMAX

Ranked attributes:

0.441	1	-0.469MMAX-0.435CHMIN-0.429CACH-0.427MMIN-0.374CHMAX...
0.302	2	0.682MYCT+0.559CHMAX-0.333MMIN+0.275CHMIN+0.152CACH...
0.179	3	0.669MYCT+0.548MMIN-0.426CHMAX+0.264MMAX-0.03CHMIN...

Selected attributes: 1,2,3 : 3