

Filling the missing values in the dataset

Step 1: select dataset "Breast-cancer"

Current relation

Relation: weather.symbolic
Instances: 14

Attributes: 5
Sum of weights: 14

Attributes

AllNoneInvertPattern

No.	Name
1	<input type="checkbox"/> outlook
2	<input checked="" type="checkbox"/> temperature
3	<input type="checkbox"/> humidity
4	<input type="checkbox"/> windy
5	<input type="checkbox"/> play

Remove

Selected attribute

Name: temperature
Missing: 0 (0%)

Distinct: 3

Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	hot	4	4
2	mild	6	6
3	cool	4	4

Class: play (Nom) Visualize All

Status
OK

Log x 0

Step 2:

Check every attribute for missing value. Here node-caps are missing values.

Current relation

Relation: breast-cancer
Instances: 286

Attributes: 10
Sum of weights: 286

Attributes

AllNoneInvertPattern

No.	Name
1	<input type="checkbox"/> age
2	<input type="checkbox"/> menopause
3	<input type="checkbox"/> tumor-size
4	<input type="checkbox"/> inv-nodes
5	<input checked="" type="checkbox"/> node-caps
6	<input type="checkbox"/> deg-malig
7	<input type="checkbox"/> breast
8	<input type="checkbox"/> breast-quad
9	<input type="checkbox"/> irradiat
10	<input type="checkbox"/> Class

Remove

Selected attribute

Name: node-caps
Missing: 8 (3%)

Distinct: 2

Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	yes	56	56
2	no	222	222

Class: Class (Nom) Visualize All

Status
OK

Log x 0

Step 3: to fill the missing value with mean/median, select filter from unsupervised -> attribute -> replace missing values and select the attribute for which the filter is to be applied and click apply

Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...

Filter
Choose
ReplaceMissingValues
Apply
Stop

Current relation
Relation: breast-cancer
Instances: 286
Attributes: 10
Sum of weights: 286

Attributes
All
None
Invert
Pattern

No.	Name
1	age
2	menopause
3	tumor-size
4	inv-nodes
5	node-caps
6	deg-malig
7	breast
8	breast-quad
9	irradiat
10	Class

Remove

Selected attribute
Name: node-caps
Missing: 8 (3%)
Distinct: 2
Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	yes	56	56
2	no	222	222

Class: Class (Nom)
Visualize All

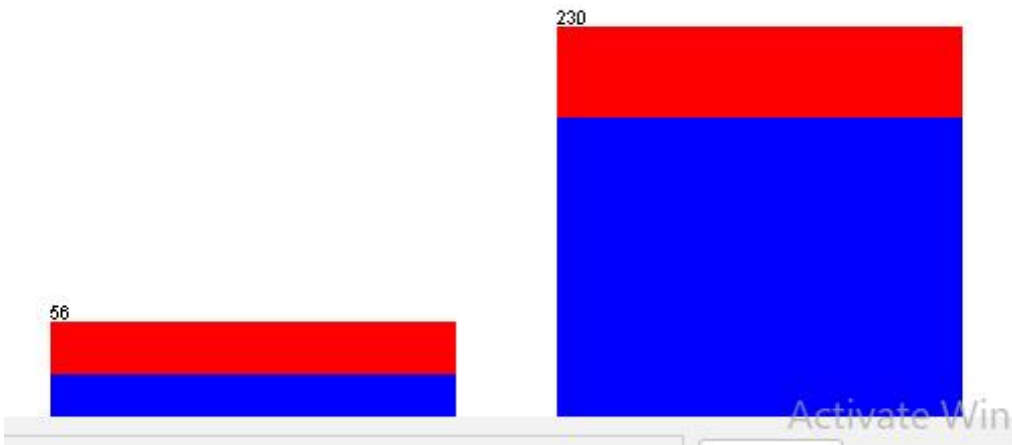
Status OK
Log
x 0

Step 4:
The missing values will be filled with mean/median.

Selected attribute
Name: node-caps
Missing: 0 (0%)
Distinct: 2
Type: Nominal
Unique: 0 (0%)

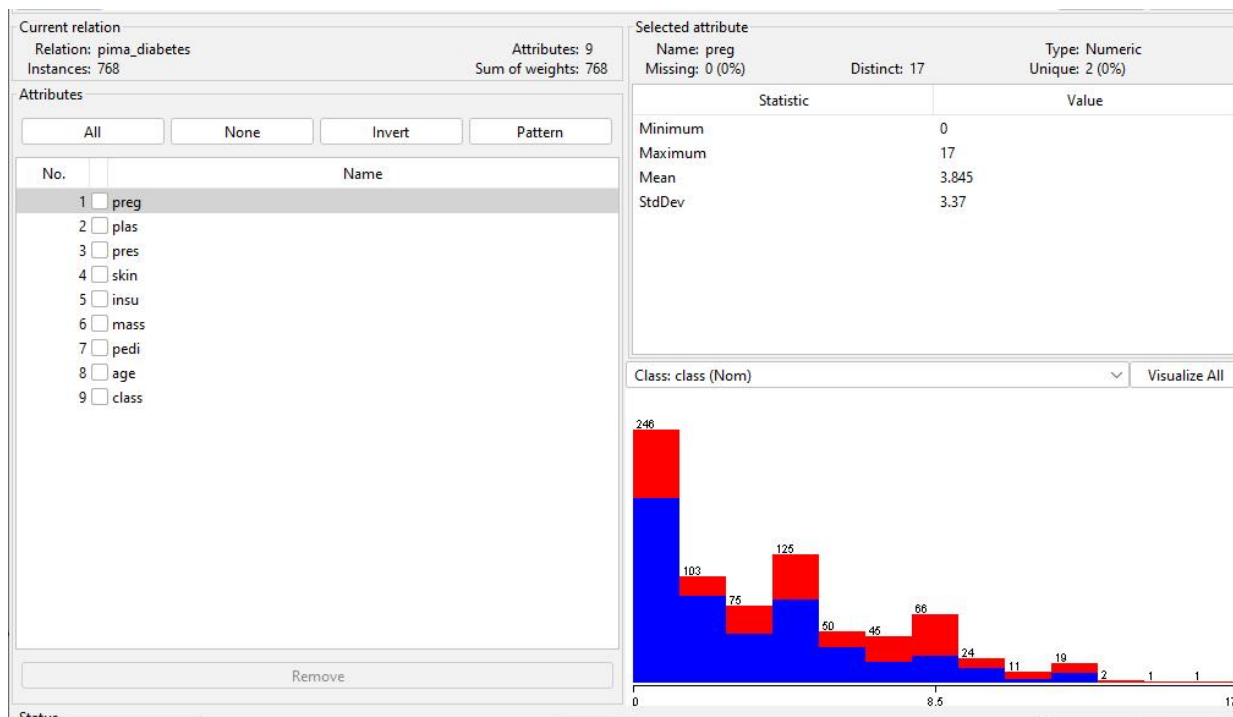
No.	Label	Count	Weight
1	yes	56	56
2	no	230	230

Class: Class (Nom)
Visualize All

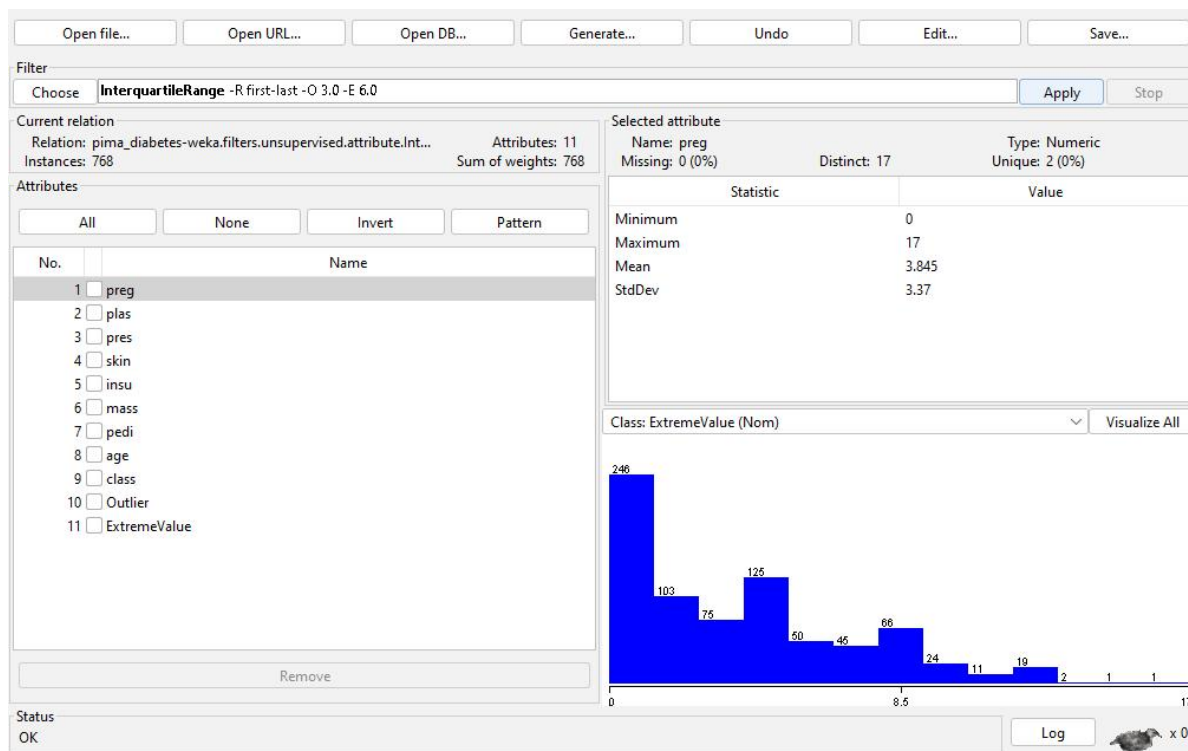


Removing outliers from the dataset

Step 1: select dataset "diabetes.arff"



Step 2: To detect the outliers, from filter section, go unsupervised -> attributes → InterquartileRange and click apply



After this, 2 more attributes are added : outlier and extreme value.

Clusterer

Choose **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

Cluster mode

☒ Use training set

☐ Supplied test set

☐ Percentage split %

☐ Classes to clusters evaluation (Nom) class

☒ Store clusters for visualization

Ignore attributes

Start

Result list (right-click for options)

04:11:09 - SimpleKMeans

Cluster output

Cluster 1: 8,95,72,0,0,36.8,0.485,57,tested_negative

Missing values globally replaced with mean/mode

Final cluster centroids:

Attribute	Full Data (768.0)	Cluster# 0 (500.0)	1 (268.0)
preg	3.8451	3.298	4.8657
plas	120.8945	109.98	141.2575
pres	69.1055	68.184	70.8246
skin	20.5365	19.664	22.1642
insu	79.7995	68.792	100.3358
mass	31.9926	30.3042	35.1425
pedi	0.4719	0.4297	0.5505
age	33.2409	31.19	37.0672
class		tested_negative	tested_positive

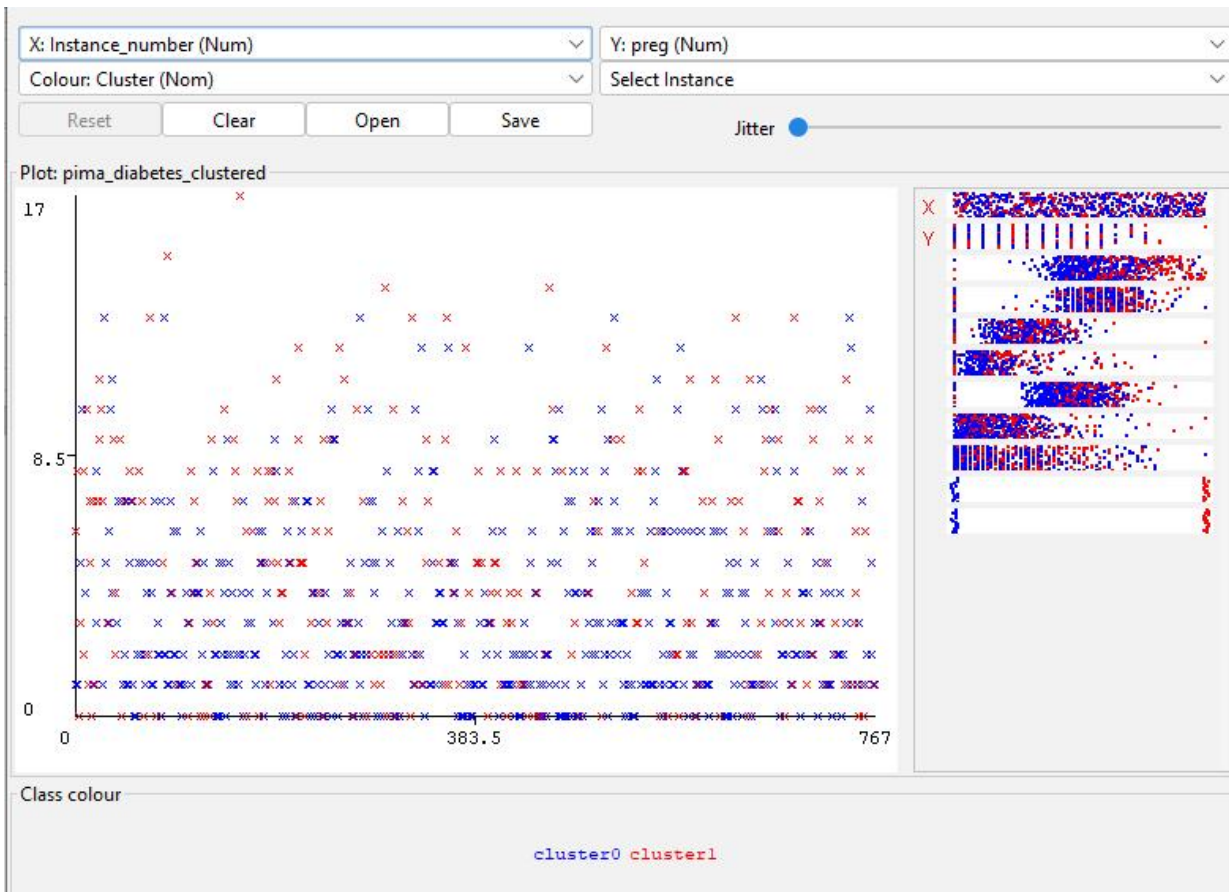
Time taken to build model (full training data) : 0.02 seconds

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

Clustered Instances

0	500 (65%)
1	268 (35%)

Step 3: To visualize, right click on the result list.



Association

Step 1: Select dataset “supermarket.arff”

Current relation
Relation: supermarket
Instances: 4627

Attributes
All None Invert Pattern

No.	Name
1	<input checked="" type="checkbox"/> department1
2	<input type="checkbox"/> department2
3	<input type="checkbox"/> department3
4	<input type="checkbox"/> department4
5	<input type="checkbox"/> department5
6	<input type="checkbox"/> department6
7	<input type="checkbox"/> department7
8	<input type="checkbox"/> department8
9	<input type="checkbox"/> department9
10	<input type="checkbox"/> grocery misc
11	<input type="checkbox"/> department11
12	<input type="checkbox"/> baby needs
13	<input type="checkbox"/> bread and cake
14	<input type="checkbox"/> baking needs
15	<input type="checkbox"/> coupons
16	<input type="checkbox"/> juice-sat-cord-ms
17	<input type="checkbox"/> tea
18	<input type="checkbox"/> biscuits

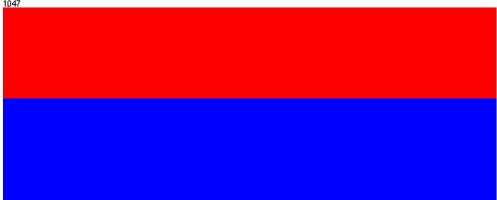
Remove

Attributes: 217
Sum of weights: 4627

Selected attribute
Name: department1
Missing: 3580 (77%)
Distinct: 1
Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	t	1047	1047

Class: total (Nom) Visualize All



Status
OK

Log x 0

Step 2: Click on the association tab and select the algorithm. We choose apriori. Right click to set the required parameters then click start.

Associator

Choose Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1

Start Stop

Result list (right-click for ...)

04:19:50 - Apriori

Associator output

Number of cycles performed: 17

Generated sets of large itemsets:

Size of set of large itemsets L(1): 44

Size of set of large itemsets L(2): 380

Size of set of large itemsets L(3): 910

Size of set of large itemsets L(4): 633

Size of set of large itemsets L(5): 105

Size of set of large itemsets L(6): 1

Best rules found:

1. biscuits=t frozen foods=t fruit=t total=high 788 ==> bread and cake=t 723 <conf:(0.92)> lift:(1.26) lev:1

2. baking needs=t biscuits=t fruit=t total=high 760 ==> bread and cake=t 696 <conf:(0.92)> lift:(1.26) lev:1

3. baking needs=t frozen foods=t fruit=t total=high 770 ==> bread and cake=t 705 <conf:(0.92)> lift:(1.26) lev:1

4. biscuits=t fruit=t vegetables=t total=high 815 ==> bread and cake=t 746 <conf:(0.92)> lift:(1.26) lev:1

5. party snack foods=t fruit=t total=high 854 ==> bread and cake=t 779 <conf:(0.91)> lift:(1.26) lev:1

6. biscuits=t frozen foods=t vegetables=t total=high 797 ==> bread and cake=t 725 <conf:(0.91)> lift:(1.26) lev:1

7. baking needs=t biscuits=t vegetables=t total=high 772 ==> bread and cake=t 701 <conf:(0.91)> lift:(1.26) lev:1

8. biscuits=t fruit=t total=high 954 ==> bread and cake=t 866 <conf:(0.91)> lift:(1.26) lev:1

9. frozen foods=t fruit=t vegetables=t total=high 834 ==> bread and cake=t 757 <conf:(0.91)> lift:(1.26) lev:1

10. frozen foods=t fruit=t total=high 969 ==> bread and cake=t 877 <conf:(0.91)> lift:(1.26) lev:1