120A3051

Shreya Idate

Batch: E3

Aim: To perform data preprocessing using WEKA

Theory:

What is data preprocessing?

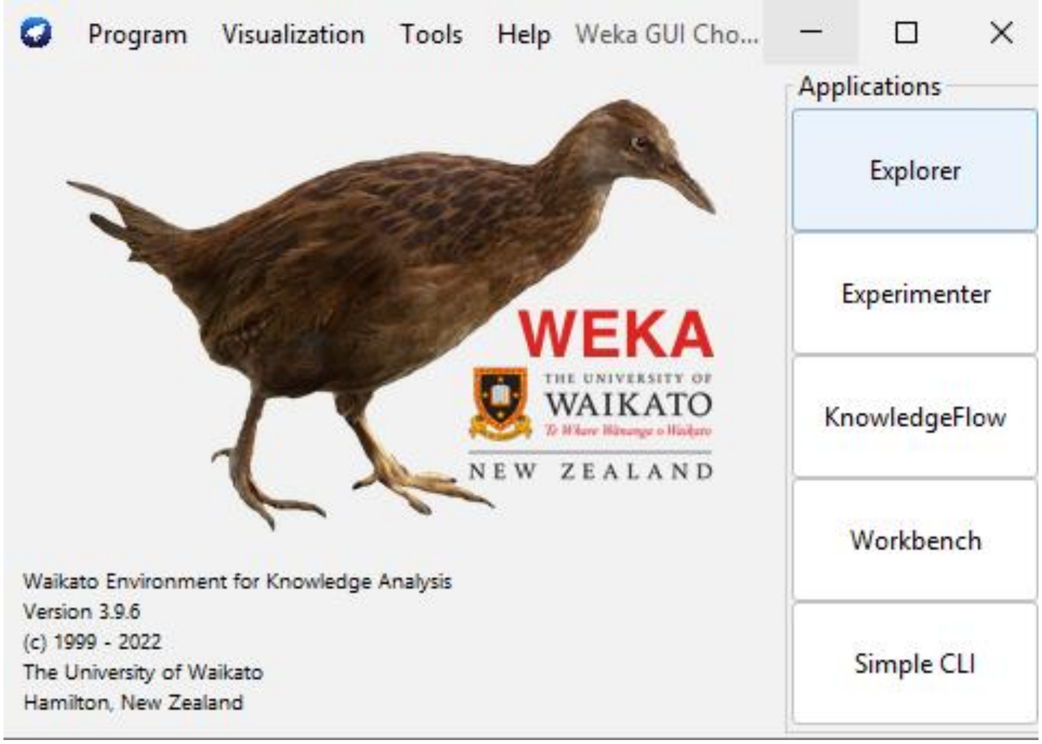
- Data preprocessing refers to the cleaning, transforming, and integrating of data in order to make it ready for analysis.
- The goal of data preprocessing is to improve the quality of the data and to make it more suitable for the specific data mining task.

Common steps in data preprocessing include:

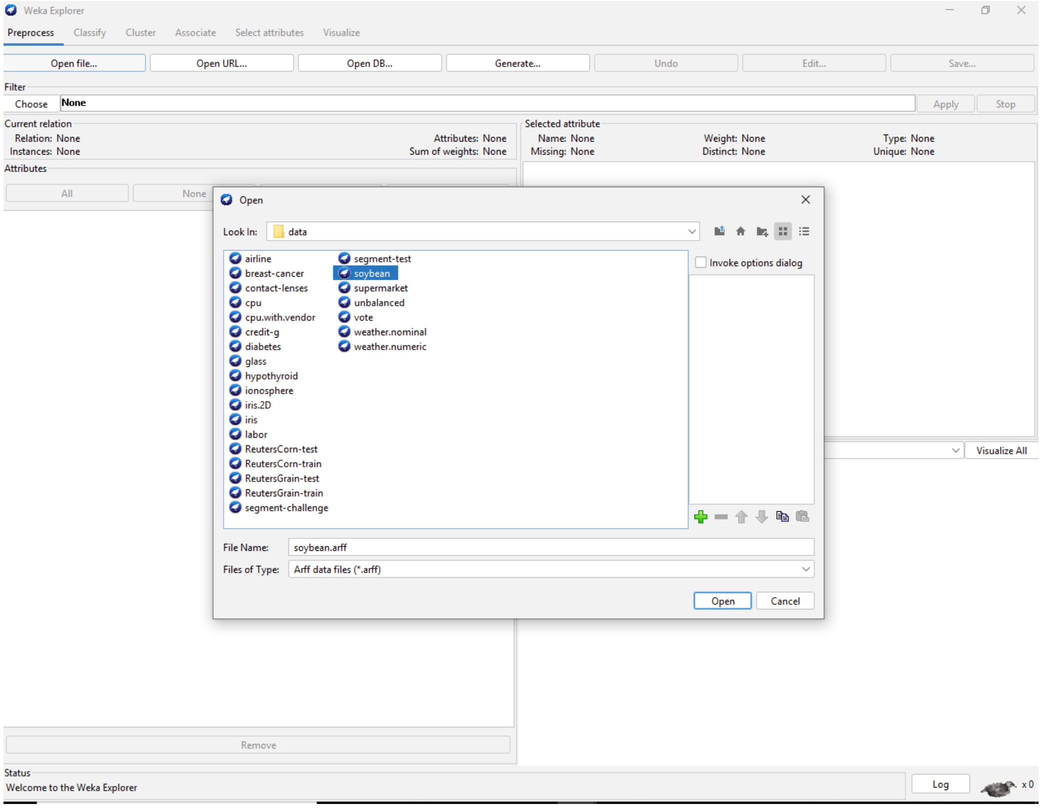
1. **Data cleaning:** this step involves identifying and removing missing, inconsistent, or irrelevant data. This can include removing duplicate records, filling in missing values, and handling outliers.
2. **Data integration:** this step involves combining data from multiple sources, such as databases, spreadsheets, and text files. The goal of integration is to create a single, consistent view of the data.
3. **Data transformation:** this step involves converting the data into a format that is more suitable for the data mining task. This can include normalizing numerical data, creating dummy variables, and encoding categorical data.
4. **Data reduction:** this step is used to select a subset of the data that is relevant to the data mining task. This can include feature selection (selecting a subset of the variables) or feature extraction (extracting new variables from the data).
5. **Data discretization:** this step is used to convert continuous numerical data into categorical data, which can be used for decision tree and other categorical data mining techniques.

Steps to perform preprocessing in WEKA:

Open Weka tool



Waikato Environment for Knowledge Analysis
Version 3.9.6
(c) 1999 - 2022
The University of Waikato
Hamilton, New Zealand



Filter: None

Relation: None
Instances: None
Attributes: None
Sum of weights: None

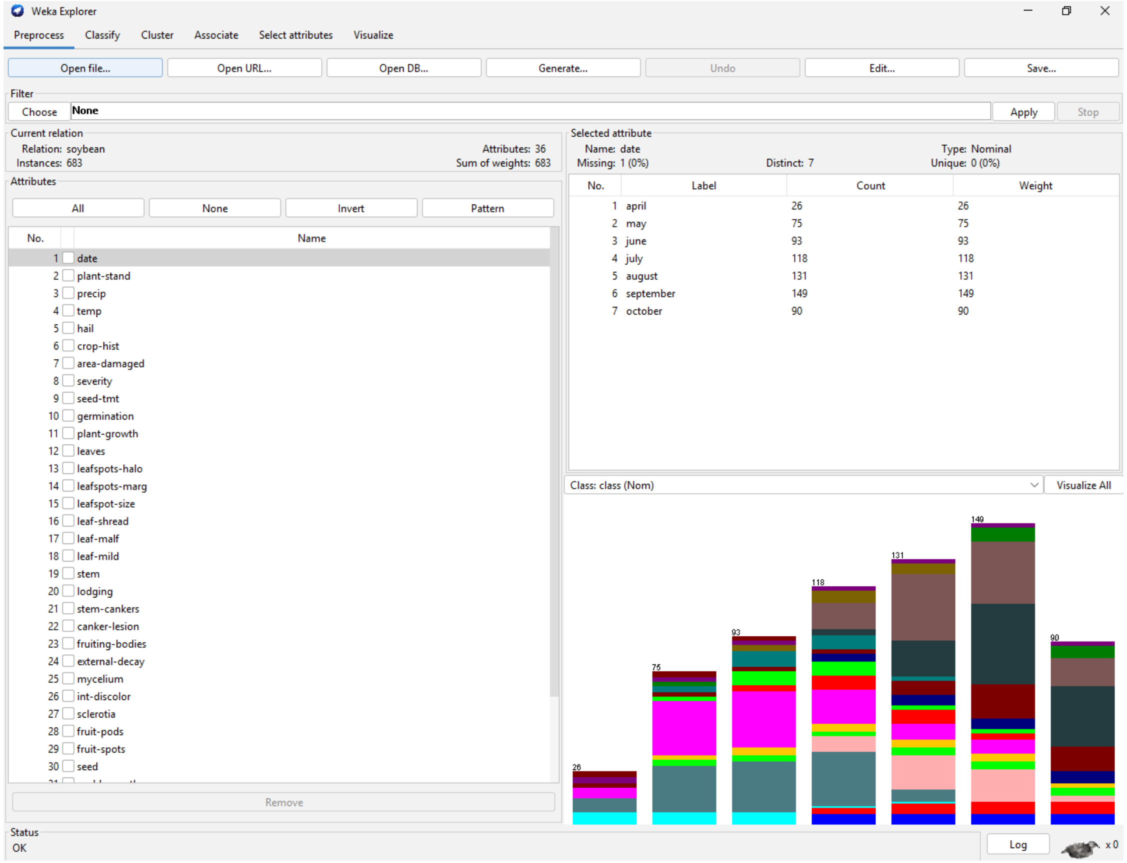
Name: None
Missing: None
Weight: None
Distinct: None
Type: None
Unique: None

Look In: data

File Name: soybean.arff

Files of Type: Arff data files (*.arff)

Status
Welcome to the Weka Explorer



Filter
Choose None

Current relation
Relation: soybean
Instances: 683
Attributes: 36
Sum of weights: 683

Attributes

No.		Name
1	☐	date
2	☐	plant-stand
3	☐	precip
4	☐	temp
5	☐	hail
6	☐	crop-hist
7	☐	area-damaged
8	☐	severity
9	☐	seed-tmt
10	☐	germination
11	☐	plant-growth
12	☐	leaves
13	☐	leafspots-halo
14	☐	leafspots-marg
15	☐	leafspot-size
16	☐	leaf-shread
17	☐	leaf-malf
18	☐	leaf-mild
19	☐	stem
20	☐	lodging
21	☐	stem-cankers
22	☐	canker-lesion
23	☐	fruiting-bodies
24	☐	external-decay
25	☐	mycelium
26	☐	int-discolor
27	☐	sclerotia
28	☐	fruit-pods
29	☐	fruit-spots
30	☐	seed

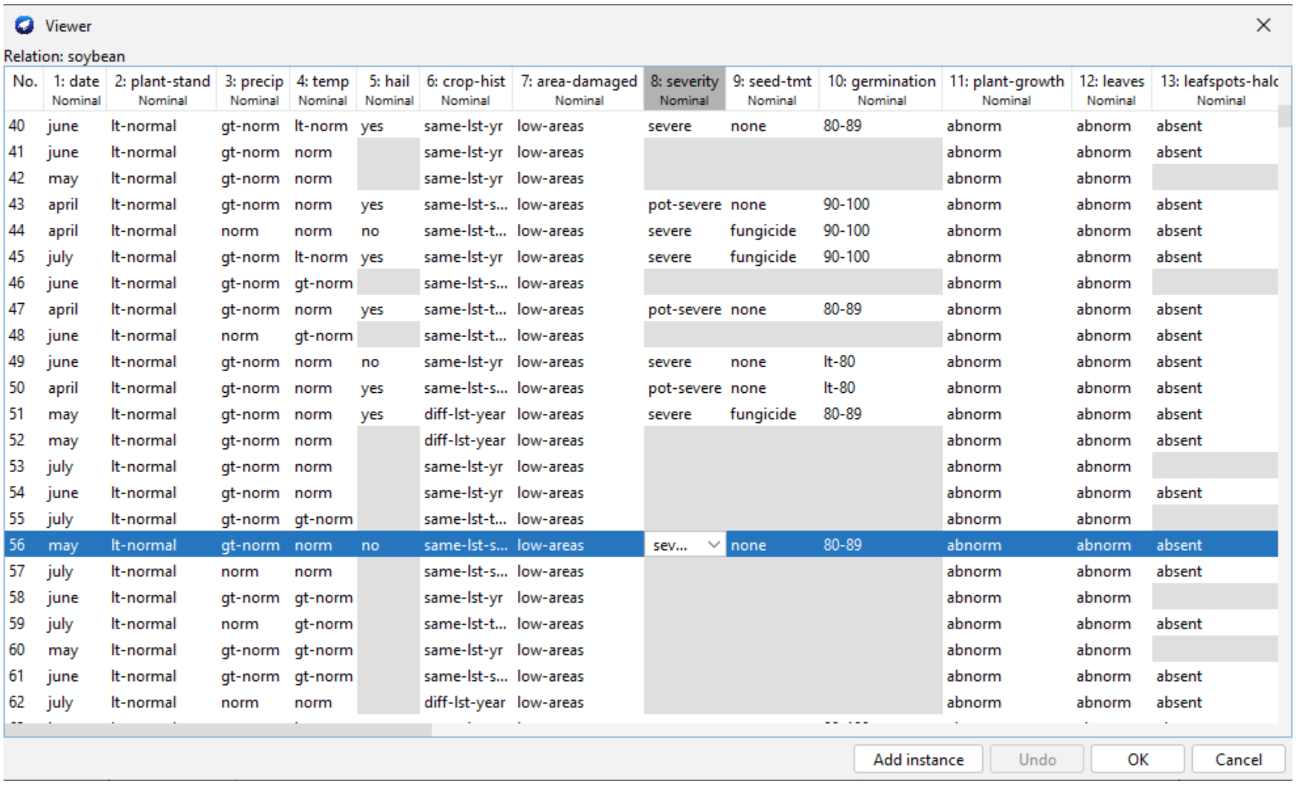
Selected attribute
Name: date
Missing: 1 (0%)
Distinct: 7
Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	april	26	26
2	may	75	75
3	june	93	93
4	july	118	118
5	august	131	131
6	september	149	149
7	october	90	90

Class: class (Nom)

Status
OK

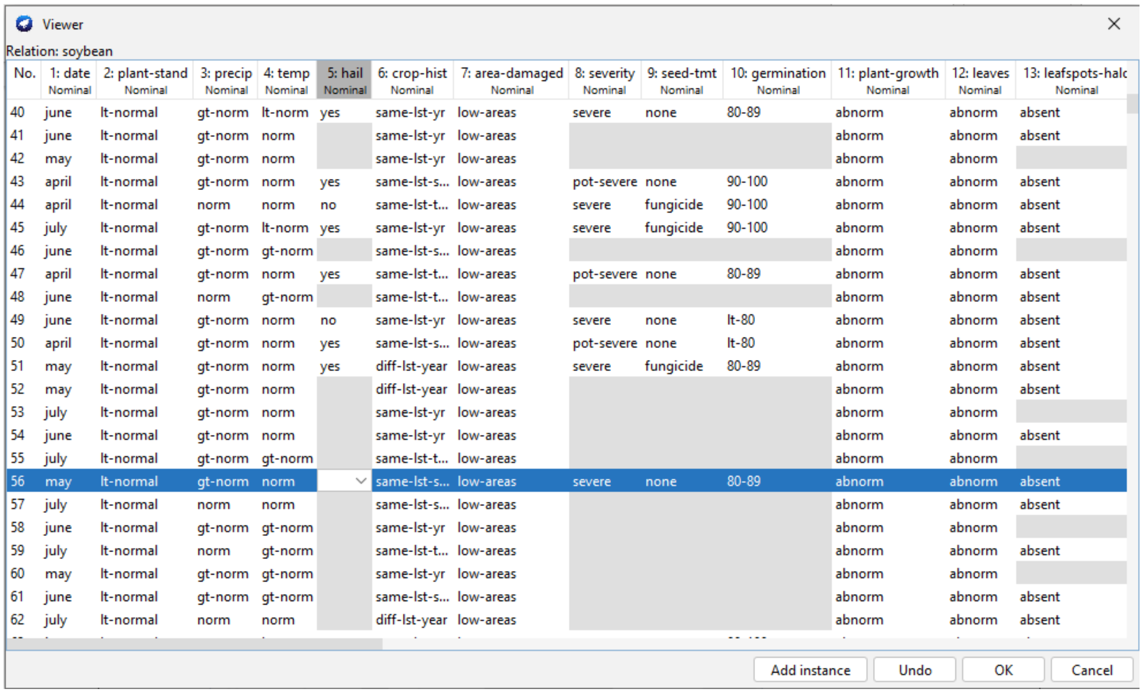
Check for missing values



Relation: soybean

No.	1: date	2: plant-stand	3: precip	4: temp	5: hail	6: crop-hist	7: area-damaged	8: severity	9: seed-tmt	10: germination	11: plant-growth	12: leaves	13: leafspots-halc
	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal
40	june	lt-normal	gt-norm	lt-norm	yes	same-lst-yr	low-areas	severe	none	80-89	abnorm	abnorm	absent
41	june	lt-normal	gt-norm	norm		same-lst-yr	low-areas				abnorm	abnorm	absent
42	may	lt-normal	gt-norm	norm		same-lst-yr	low-areas				abnorm	abnorm	
43	april	lt-normal	gt-norm	norm	yes	same-lst-s...	low-areas	pot-severe	none	90-100	abnorm	abnorm	absent
44	april	lt-normal	norm	norm	no	same-lst-t...	low-areas	severe	fungicide	90-100	abnorm	abnorm	absent
45	july	lt-normal	gt-norm	lt-norm	yes	same-lst-yr	low-areas	severe	fungicide	90-100	abnorm	abnorm	absent
46	june	lt-normal	gt-norm	gt-norm		same-lst-s...	low-areas				abnorm	abnorm	
47	april	lt-normal	gt-norm	norm	yes	same-lst-t...	low-areas	pot-severe	none	80-89	abnorm	abnorm	absent
48	june	lt-normal	norm	gt-norm		same-lst-t...	low-areas				abnorm	abnorm	absent
49	june	lt-normal	gt-norm	norm	no	same-lst-yr	low-areas	severe	none	lt-80	abnorm	abnorm	absent
50	april	lt-normal	gt-norm	norm	yes	same-lst-s...	low-areas	pot-severe	none	lt-80	abnorm	abnorm	absent
51	may	lt-normal	gt-norm	norm	yes	diff-lst-year	low-areas	severe	fungicide	80-89	abnorm	abnorm	absent
52	may	lt-normal	gt-norm	norm		diff-lst-year	low-areas				abnorm	abnorm	absent
53	july	lt-normal	gt-norm	norm		same-lst-yr	low-areas				abnorm	abnorm	
54	june	lt-normal	gt-norm	norm		same-lst-yr	low-areas				abnorm	abnorm	absent
55	july	lt-normal	gt-norm	gt-norm		same-lst-t...	low-areas				abnorm	abnorm	
56	may	lt-normal	gt-norm	norm	no	same-lst-s...	low-areas	sev...	none	80-89	abnorm	abnorm	absent
57	july	lt-normal	norm	norm		same-lst-s...	low-areas				abnorm	abnorm	absent
58	june	lt-normal	gt-norm	gt-norm		same-lst-yr	low-areas				abnorm	abnorm	
59	july	lt-normal	norm	gt-norm		same-lst-t...	low-areas				abnorm	abnorm	absent
60	may	lt-normal	gt-norm	gt-norm		same-lst-yr	low-areas				abnorm	abnorm	
61	june	lt-normal	gt-norm	gt-norm		same-lst-s...	low-areas				abnorm	abnorm	absent
62	july	lt-normal	norm	norm		diff-lst-year	low-areas				abnorm	abnorm	absent

Create missing values by selecting any one and deleting data

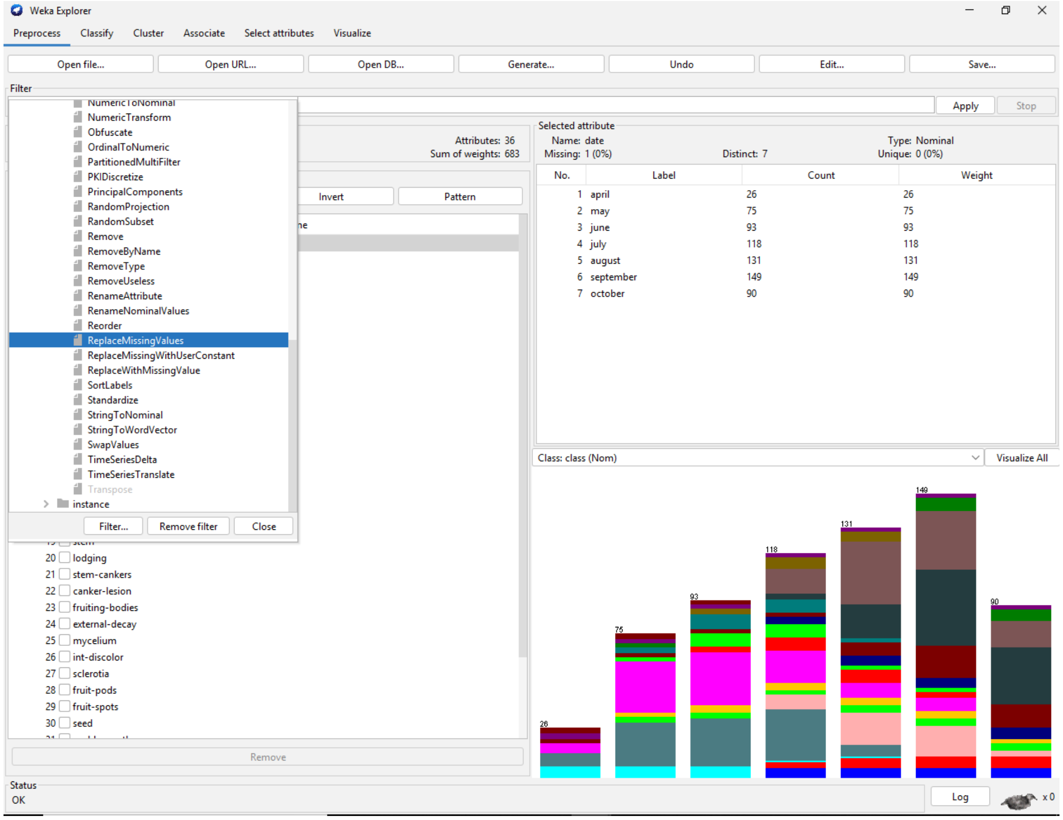


Relation: soybean

No.	1: date Nominal	2: plant-stand Nominal	3: precip Nominal	4: temp Nominal	5: hail Nominal	6: crop-hist Nominal	7: area-damaged Nominal	8: severity Nominal	9: seed-tmt Nominal	10: germination Nominal	11: plant-growth Nominal	12: leaves Nominal	13: leafspots-halc Nominal
40	june	lt-normal	gt-norm	lt-norm	yes	same-lst-yr	low-areas	severe	none	80-89	abnorm	abnorm	absent
41	june	lt-normal	gt-norm	norm		same-lst-yr	low-areas				abnorm	abnorm	absent
42	may	lt-normal	gt-norm	norm		same-lst-yr	low-areas				abnorm	abnorm	
43	april	lt-normal	gt-norm	norm	yes	same-lst-s...	low-areas	pot-severe	none	90-100	abnorm	abnorm	absent
44	april	lt-normal	norm	norm	no	same-lst-t...	low-areas	severe	fungicide	90-100	abnorm	abnorm	absent
45	july	lt-normal	gt-norm	lt-norm	yes	same-lst-yr	low-areas	severe	fungicide	90-100	abnorm	abnorm	absent
46	june	lt-normal	gt-norm	gt-norm		same-lst-s...	low-areas				abnorm	abnorm	
47	april	lt-normal	gt-norm	norm	yes	same-lst-t...	low-areas	pot-severe	none	80-89	abnorm	abnorm	absent
48	june	lt-normal	norm	gt-norm		same-lst-t...	low-areas				abnorm	abnorm	absent
49	june	lt-normal	gt-norm	norm	no	same-lst-yr	low-areas	severe	none	lt-80	abnorm	abnorm	absent
50	april	lt-normal	gt-norm	norm	yes	same-lst-s...	low-areas	pot-severe	none	lt-80	abnorm	abnorm	absent
51	may	lt-normal	gt-norm	norm	yes	diff-lst-year	low-areas	severe	fungicide	80-89	abnorm	abnorm	absent
52	may	lt-normal	gt-norm	norm		diff-lst-year	low-areas				abnorm	abnorm	absent
53	july	lt-normal	gt-norm	norm		same-lst-yr	low-areas				abnorm	abnorm	
54	june	lt-normal	gt-norm	norm		same-lst-yr	low-areas				abnorm	abnorm	absent
55	july	lt-normal	gt-norm	gt-norm		same-lst-t...	low-areas				abnorm	abnorm	
56	may	lt-normal	gt-norm	norm		same-lst-s...	low-areas	severe	none	80-89	abnorm	abnorm	absent
57	july	lt-normal	norm	norm		same-lst-s...	low-areas				abnorm	abnorm	absent
58	june	lt-normal	gt-norm	gt-norm		same-lst-yr	low-areas				abnorm	abnorm	
59	july	lt-normal	norm	gt-norm		same-lst-t...	low-areas				abnorm	abnorm	absent
60	may	lt-normal	gt-norm	gt-norm		same-lst-yr	low-areas				abnorm	abnorm	
61	june	lt-normal	gt-norm	gt-norm		same-lst-s...	low-areas				abnorm	abnorm	absent
62	july	lt-normal	norm	norm		diff-lst-year	low-areas				abnorm	abnorm	absent

Choose a filter to apply:

1) Using replace missing values



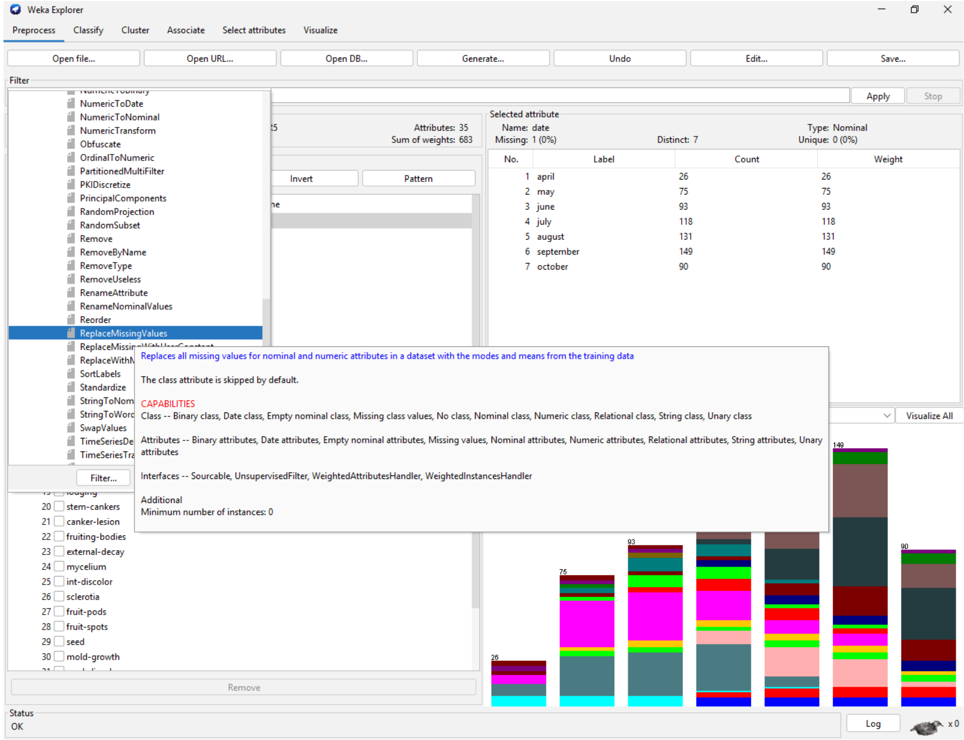
Attributes: 36
Sum of weights: 683

20 □ lodging
21 □ stem-cankers
22 □ canker-lesion
23 □ fruiting-bodies
24 □ external-decay
25 □ mycelium
26 □ int-discolor
27 □ sclerotia
28 □ fruit-pods
29 □ fruit-spots
30 □ seed

Selected attribute
Name: date
Missing: 1 (0%)
Distinct: 7
Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	april	26	26
2	may	75	75
3	june	93	93
4	july	118	118
5	august	131	131
6	september	149	149
7	october	90	90

Status
OK



Attributes: 35
Sum of weights: 683

Selected attribute
Name: date
Type: Nominal
Missing: 1 (0%)
Distinct: 7
Unique: 0 (0%)

No.	Label	Count	Weight
1	april	26	26
2	may	75	75
3	june	93	93
4	july	118	118
5	august	131	131
6	september	149	149
7	october	90	90

Replaces all missing values for nominal and numeric attributes in a dataset with the modes and means from the training data

The class attribute is skipped by default.

CAPABILITIES
Class -- Binary class, Date class, Empty nominal class, Missing class values, No class, Nominal class, Numeric class, Relational class, String class, Unary class

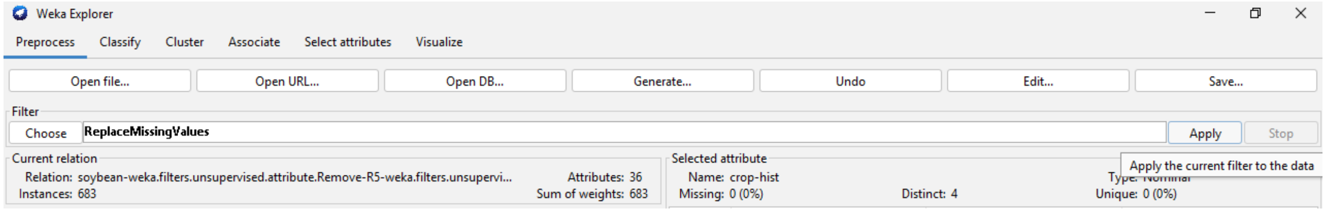
Attributes -- Binary attributes, Date attributes, Empty nominal attributes, Missing values, Nominal attributes, Numeric attributes, Relational attributes, String attributes, Unary attributes

Interfaces -- Sourcable, UnsupervisedFilter, WeightedAttributesHandler, WeightedInstancesHandler

Additional
Minimum number of instances: 0

Status
OK

Click Apply to apply the filter to the data

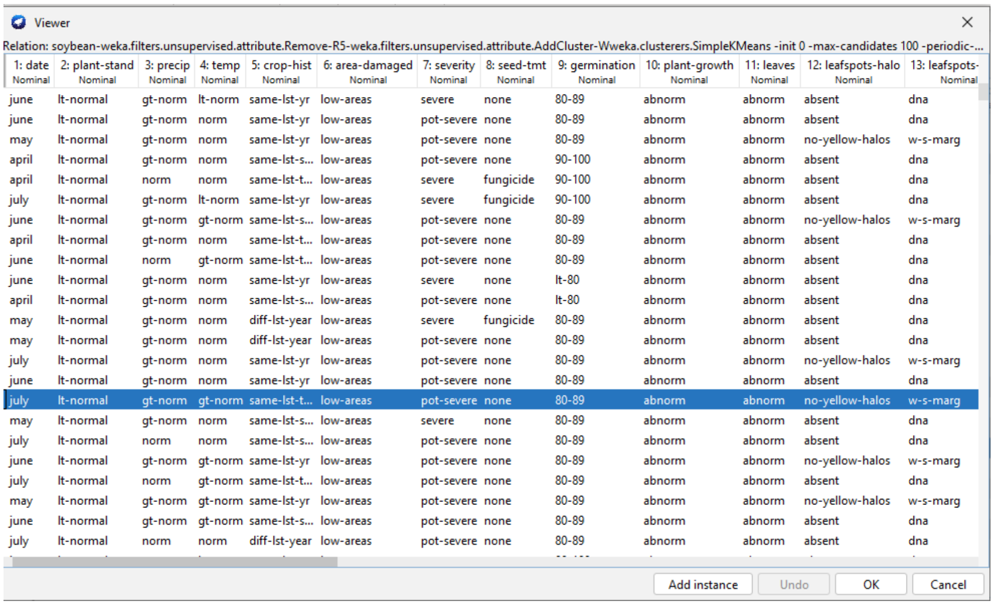


ReplaceMissingValues

Current relation
Relation: soybean-weka.filters.unsupervised.attribute.Remove-R5-weka.filters.unsupervi...
Instances: 683
Attributes: 36
Sum of weights: 683

Selected attribute
Name: crop-hist
Type: Nominal
Missing: 0 (0%)
Distinct: 4
Unique: 0 (0%)

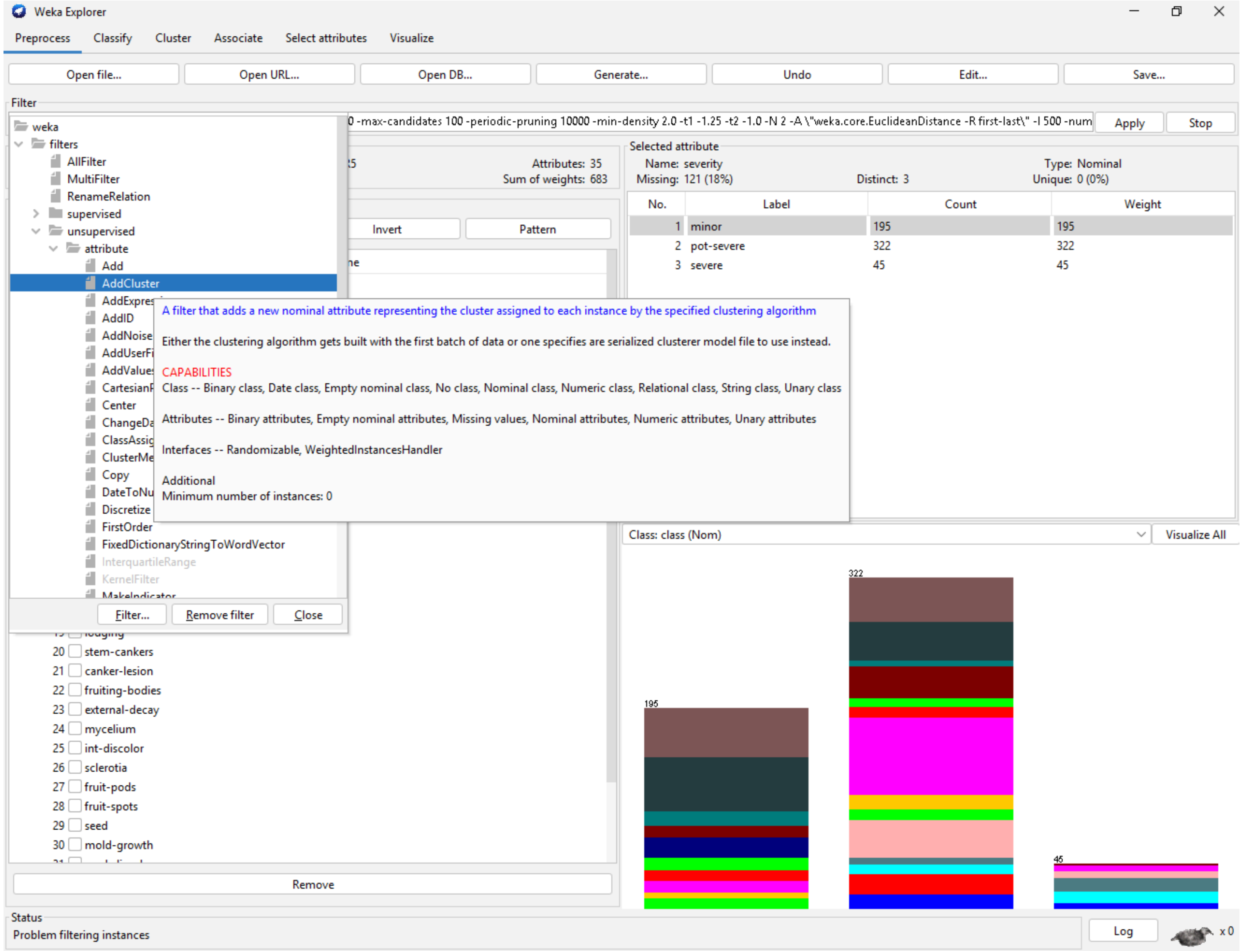
All missing values are now replaced



Relation: soybean-weka.filters.unsupervised.attribute.Remove-R5-weka.filters.unsupervised.attribute.AddCluster-Wweka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-...

1: date Nominal	2: plant-stand Nominal	3: precip Nominal	4: temp Nominal	5: crop-hist Nominal	6: area-damaged Nominal	7: severity Nominal	8: seed-tmt Nominal	9: germination Nominal	10: plant-growth Nominal	11: leaves Nominal	12: leafspots-halo Nominal	13: leafspots- Nominal
june	lt-normal	gt-norm	lt-norm	same-lst-yr	low-areas	severe	none	80-89	abnorm	abnorm	absent	dna
june	lt-normal	gt-norm	norm	same-lst-yr	low-areas	pot-severe	none	80-89	abnorm	abnorm	absent	dna
may	lt-normal	gt-norm	norm	same-lst-yr	low-areas	pot-severe	none	80-89	abnorm	abnorm	no-yellow-halos	w-s-marg
april	lt-normal	gt-norm	norm	same-lst-s...	low-areas	pot-severe	none	90-100	abnorm	abnorm	absent	dna
april	lt-normal	norm	norm	same-lst-t...	low-areas	severe	fungicide	90-100	abnorm	abnorm	absent	dna
july	lt-normal	gt-norm	lt-norm	same-lst-yr	low-areas	severe	fungicide	90-100	abnorm	abnorm	absent	dna
june	lt-normal	gt-norm	gt-norm	same-lst-s...	low-areas	pot-severe	none	80-89	abnorm	abnorm	no-yellow-halos	w-s-marg
april	lt-normal	gt-norm	norm	same-lst-t...	low-areas	pot-severe	none	80-89	abnorm	abnorm	absent	dna
june	lt-normal	norm	gt-norm	same-lst-t...	low-areas	pot-severe	none	80-89	abnorm	abnorm	absent	dna
june	lt-normal	gt-norm	norm	same-lst-yr	low-areas	severe	none	lt-80	abnorm	abnorm	absent	dna
april	lt-normal	gt-norm	norm	same-lst-s...	low-areas	pot-severe	none	lt-80	abnorm	abnorm	absent	dna
may	lt-normal	gt-norm	norm	diff-lst-year	low-areas	severe	fungicide	80-89	abnorm	abnorm	absent	dna
may	lt-normal	gt-norm	norm	diff-lst-year	low-areas	pot-severe	none	80-89	abnorm	abnorm	absent	dna
july	lt-normal	gt-norm	norm	same-lst-yr	low-areas	pot-severe	none	80-89	abnorm	abnorm	no-yellow-halos	w-s-marg
june	lt-normal	gt-norm	norm	same-lst-yr	low-areas	pot-severe	none	80-89	abnorm	abnorm	absent	dna
july	lt-normal	gt-norm	gt-norm	same-lst-t...	low-areas	pot-severe	none	80-89	abnorm	abnorm	no-yellow-halos	w-s-marg
may	lt-normal	gt-norm	norm	same-lst-s...	low-areas	severe	none	80-89	abnorm	abnorm	absent	dna
july	lt-normal	norm	norm	same-lst-s...	low-areas	pot-severe	none	80-89	abnorm	abnorm	absent	dna
june	lt-normal	gt-norm	gt-norm	same-lst-yr	low-areas	pot-severe	none	80-89	abnorm	abnorm	no-yellow-halos	w-s-marg
july	lt-normal	norm	gt-norm	same-lst-t...	low-areas	pot-severe	none	80-89	abnorm	abnorm	absent	dna
may	lt-normal	gt-norm	gt-norm	same-lst-yr	low-areas	pot-severe	none	80-89	abnorm	abnorm	no-yellow-halos	w-s-marg
june	lt-normal	gt-norm	gt-norm	same-lst-s...	low-areas	pot-severe	none	80-89	abnorm	abnorm	absent	dna
july	lt-normal	norm	norm	diff-lst-year	low-areas	pot-severe	none	80-89	abnorm	abnorm	absent	dna

2) Using Add Cluster filter



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 500 -num

- weka
  - filters
    - AllFilter
    - MultiFilter
    - RenameRelation
    - supervised
    - unsupervised
      - attribute
        - Add
        - AddCluster
        - AddExpres
        - AddID
        - AddNoise
        - AddUserFi
        - AddValue
        - CartesianP
        - Center
        - ChangeDa
        - ClassAssig
        - ClusterMe
        - Copy
        - DateToNu
        - Discretize
        - FirstOrder
        - FixedDictionaryStringToWordVector
        - InterquartileRange
        - KernelFilter
        - MakeIndicator

A filter that adds a new nominal attribute representing the cluster assigned to each instance by the specified clustering algorithm

Either the clustering algorithm gets built with the first batch of data or one specifies are serialized clusterer model file to use instead.

CAPABILITIES
Class -- Binary class, Date class, Empty nominal class, No class, Nominal class, Numeric class, Relational class, String class, Unary class

Attributes -- Binary attributes, Empty nominal attributes, Missing values, Nominal attributes, Numeric attributes, Unary attributes

Interfaces -- Randomizable, WeightedInstancesHandler

Additional
Minimum number of instances: 0

Attributes: 35
Sum of weights: 683

Selected attribute
Name: severity
Missing: 121 (18%)
Distinct: 3
Type: Nominal
Unique: 0 (0%)

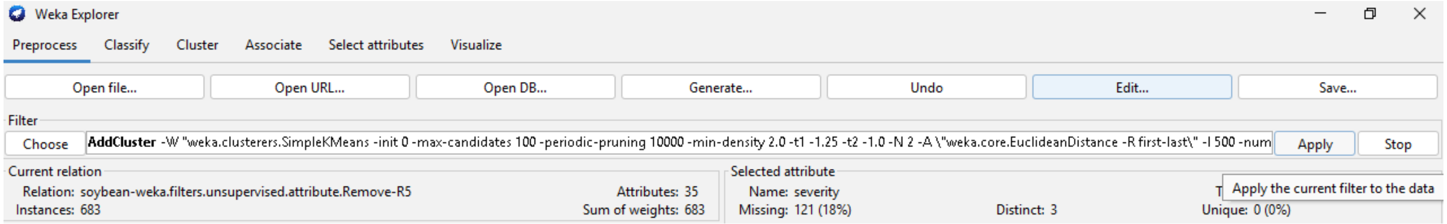
No.	Label	Count	Weight
1	minor	195	195
2	pot-severe	322	322
3	severe	45	45

Class: class (Nom)

20	stem-cankers
21	canker-lesion
22	fruiting-bodies
23	external-decay
24	mycelium
25	int-discolor
26	sclerotia
27	fruit-pods
28	fruit-spots
29	seed
30	mold-growth

Status
Problem filtering instances

Click Apply to apply the filter to the data

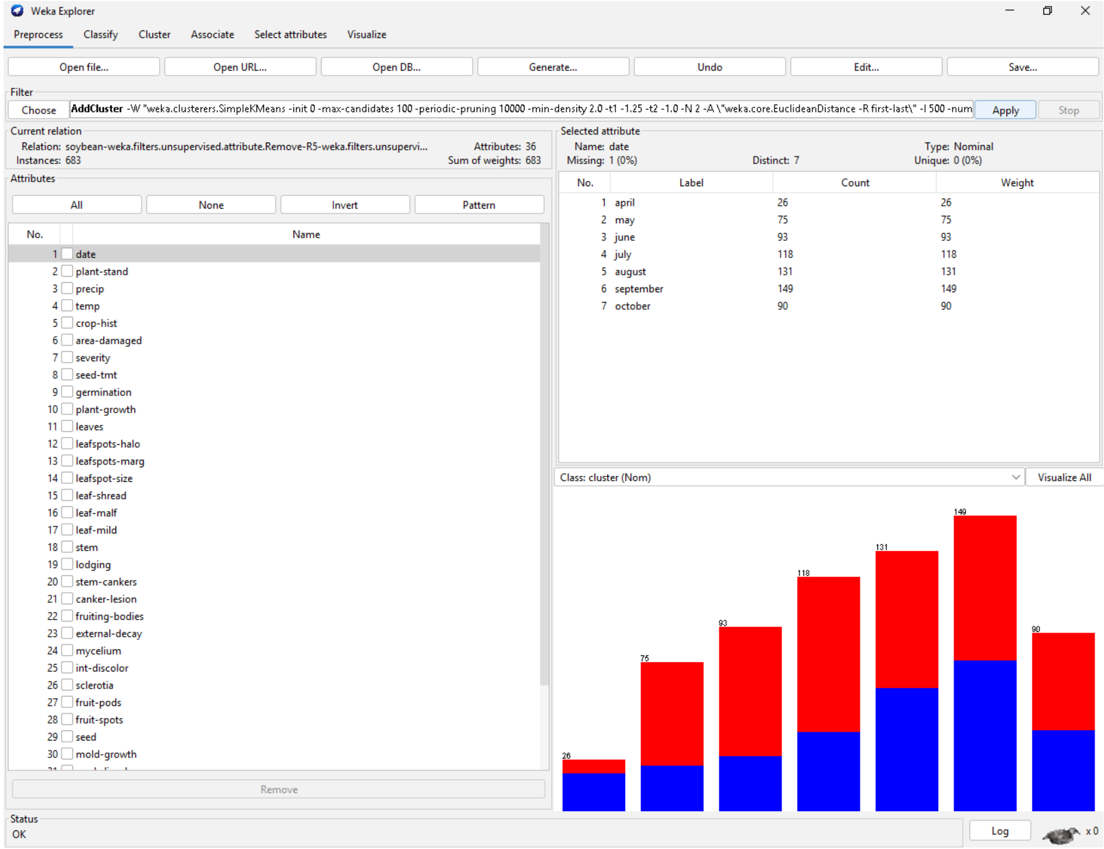


AddCluster -W "weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A \"weka.core.EuclideanDistance -R first-last\" -I 500 -num

Current relation
Relation: soybean-weka.filters.unsupervised.attribute.Remove-R5
Instances: 683
Attributes: 35
Sum of weights: 683

Selected attribute
Name: severity
Missing: 121 (18%)
Distinct: 3
T
Unique: 0 (0%)

Apply the current filter to the data



Weka Explorer

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

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

Filter

Choose **AddCluster** -W "weka.clusterers.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 500 -num

Apply | Stop

Current relation

Relation: soybean-weka.filters.unsupervised.attribute.Remove-R5-weka.filters.unsupervi... Attributes: 36
Instances: 683 Sum of weights: 683

Selected attribute

Name: date Type: Nominal
Missing: 1 (0%) Distinct: 7 Unique: 0 (0%)

No.	Label	Count	Weight
1	april	26	26
2	may	75	75
3	june	93	93
4	july	118	118
5	august	131	131
6	september	149	149
7	october	90	90

Attributes

All | None | Invert | Pattern

No.		Name
1	□	date
2	□	plant-stand
3	□	precip
4	□	temp
5	□	crop-hist
6	□	area-damaged
7	□	severity
8	□	seed-tmt
9	□	germination
10	□	plant-growth
11	□	leaves
12	□	leafspots-halo
13	□	leafspots-marg
14	□	leafspot-size
15	□	leaf-shread
16	□	leaf-malf
17	□	leaf-mild
18	□	stem
19	□	lodging
20	□	stem-cankers
21	□	canker-lesion
22	□	fruiting-bodies
23	□	external-decay
24	□	mycelium
25	□	int-discolor
26	□	sclerotia
27	□	fruit-pods
28	□	fruit-spots
29	□	seed
30	□	mold-growth

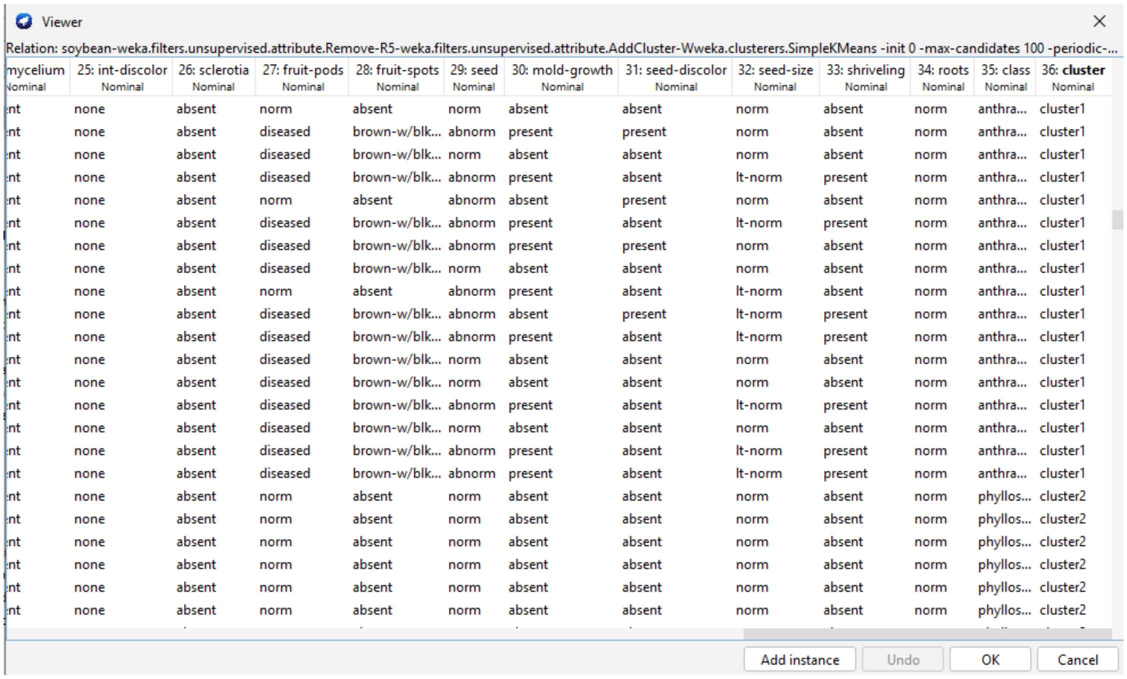
Remove

Class: cluster (Nom) | Visualize All

Status
OK

Log | x 0

Click edit to view the last column created when clusters were created



Viewer

Relation: soybean-weka.filters.unsupervised.attribute.Remove-R5-weka.filters.unsupervised.attribute.AddCluster-Wweka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-...

mycelium Nominal	25: int-discolor Nominal	26: sclerotia Nominal	27: fruit-pods Nominal	28: fruit-spots Nominal	29: seed Nominal	30: mold-growth Nominal	31: seed-discolor Nominal	32: seed-size Nominal	33: shriveling Nominal	34: roots Nominal	35: class Nominal	36: **cluster** Nominal
nt	none	absent	norm	absent	norm	absent	absent	norm	absent	norm	anthra...	cluster1
nt	none	absent	diseased	brown-w/blk...	abnorm	present	present	norm	absent	norm	anthra...	cluster1
nt	none	absent	diseased	brown-w/blk...	norm	absent	absent	norm	absent	norm	anthra...	cluster1
nt	none	absent	diseased	brown-w/blk...	abnorm	present	absent	lt-norm	present	norm	anthra...	cluster1
nt	none	absent	norm	absent	abnorm	absent	present	norm	absent	norm	anthra...	cluster1
nt	none	absent	diseased	brown-w/blk...	abnorm	present	absent	lt-norm	present	norm	anthra...	cluster1
nt	none	absent	diseased	brown-w/blk...	abnorm	present	present	norm	absent	norm	anthra...	cluster1
nt	none	absent	diseased	brown-w/blk...	norm	absent	absent	norm	absent	norm	anthra...	cluster1
nt	none	absent	norm	absent	abnorm	present	absent	lt-norm	absent	norm	anthra...	cluster1
nt	none	absent	diseased	brown-w/blk...	abnorm	absent	present	lt-norm	present	norm	anthra...	cluster1
nt	none	absent	diseased	brown-w/blk...	abnorm	present	absent	lt-norm	present	norm	anthra...	cluster1
nt	none	absent	diseased	brown-w/blk...	norm	absent	absent	norm	absent	norm	anthra...	cluster1
nt	none	absent	diseased	brown-w/blk...	norm	absent	absent	norm	absent	norm	anthra...	cluster1
nt	none	absent	diseased	brown-w/blk...	abnorm	present	absent	lt-norm	present	norm	anthra...	cluster1
nt	none	absent	diseased	brown-w/blk...	norm	absent	absent	norm	absent	norm	anthra...	cluster1
nt	none	absent	diseased	brown-w/blk...	abnorm	present	absent	lt-norm	present	norm	anthra...	cluster1
nt	none	absent	diseased	brown-w/blk...	abnorm	present	absent	lt-norm	present	norm	anthra...	cluster1
nt	none	absent	norm	absent	norm	absent	absent	norm	absent	norm	phyllos...	cluster2
nt	none	absent	norm	absent	norm	absent	absent	norm	absent	norm	phyllos...	cluster2
nt	none	absent	norm	absent	norm	absent	absent	norm	absent	norm	phyllos...	cluster2
nt	none	absent	norm	absent	norm	absent	absent	norm	absent	norm	phyllos...	cluster2
nt	none	absent	norm	absent	norm	absent	absent	norm	absent	norm	phyllos...	cluster2
nt	none	absent	norm	absent	norm	absent	absent	norm	absent	norm	phyllos...	cluster2

Add instance | Undo | OK | Cancel

Conclusion: Successfully performed data preprocessing using WEKA.