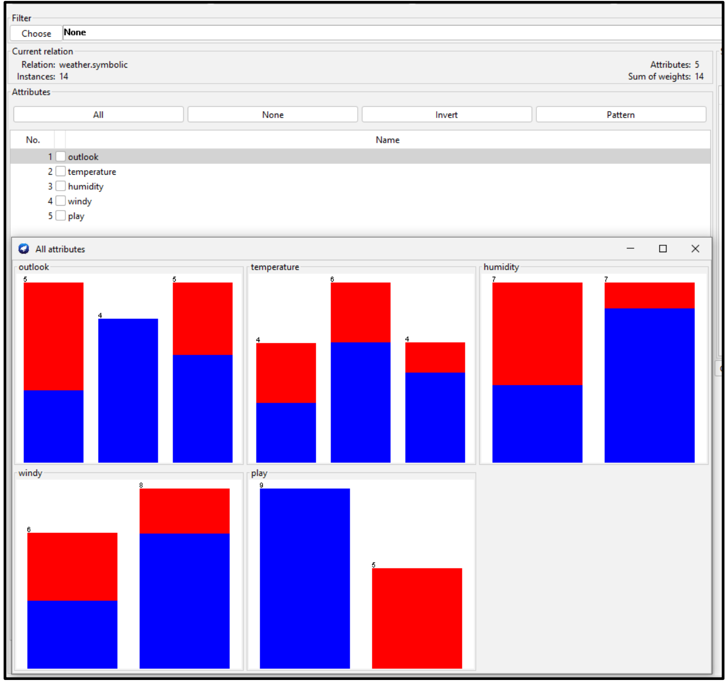
Experiment 4

Title: Apply Preprocessing techniques on dataset using filters: Remove, ReplaceMissingValues, ReplaceMissingWithUserConstant, ReplaceWithMissingValue, Descritize. Also do the result analysis before and after preprocessing.

1. Remove filter: The Remove filter is an unsupervised attribute filter that allows users to delete specific columns from a dataset. It is particularly useful when you want to exclude irrelevant, redundant or sensitive attributes from the data before applying machine learning algorithms.

Before applying the "Remove filter"

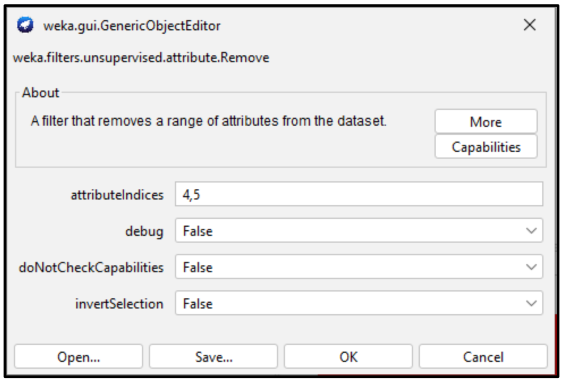


Filter: None

Current relation
Relation: weather.symbolic
Instances: 14
Attributes: 5
Sum of weights: 14

Attributes

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



weka.gui.GenericObjectEditor

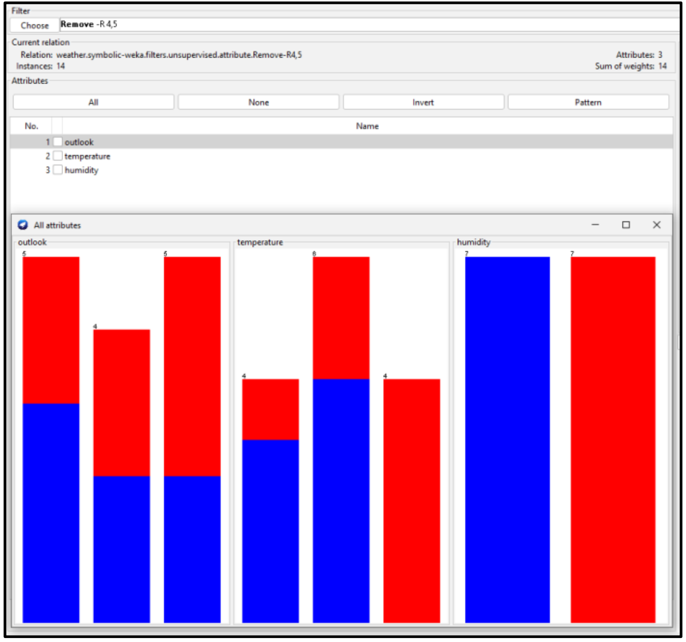
weka.filters.unsupervised.attribute.Remove

About

A filter that removes a range of attributes from the dataset.

attributeIndices: 4,5
debug: False
doNotCheckCapabilities: False
invertSelection: False

after applying the “Remove filter”



Remove -R 4,5

Current relation

Relation: weather.symbolic-weka.filters.unsupervised.attribute.Remove-R4,5

Instances: 14

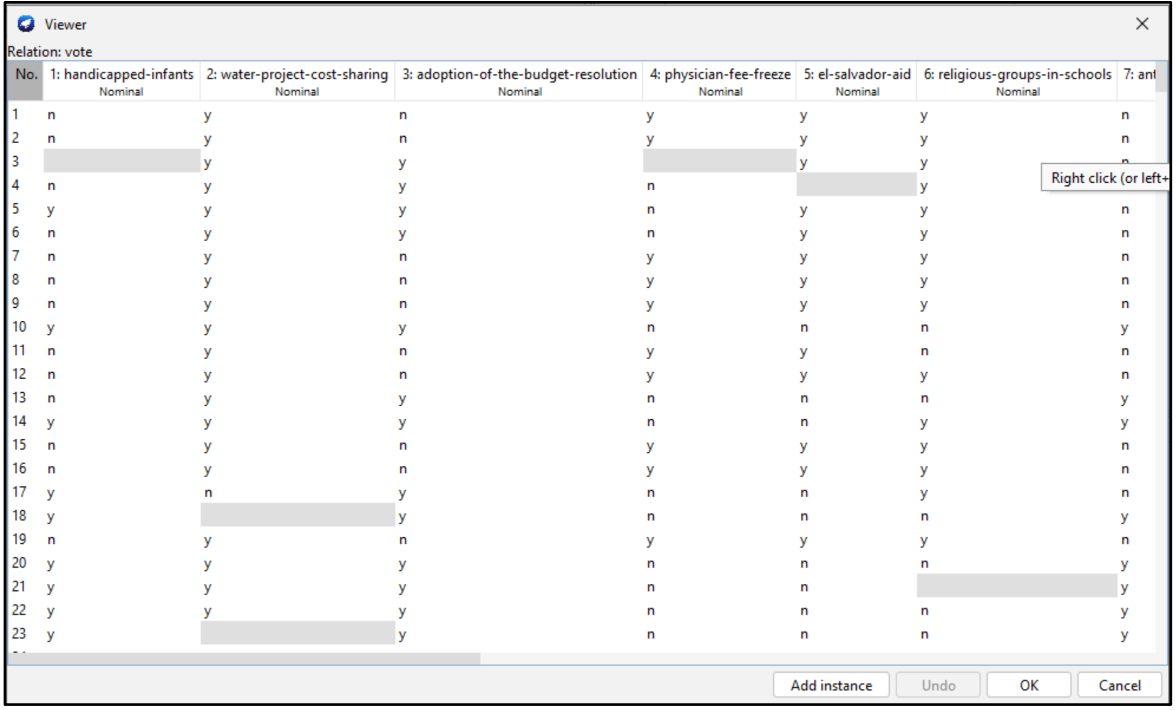
Attributes: 3

Sum of weights: 14

No.		Name
1	□	outlook
2	□	temperature
3	□	humidity

2. ReplaceMissingValues filter: The ReplaceMissingValues filter is an unsupervised attribute and instance filter used to automatically fill in missing values in dataset. It ensures that incomplete data does not negatively impact the performance of machine learning algorithms.

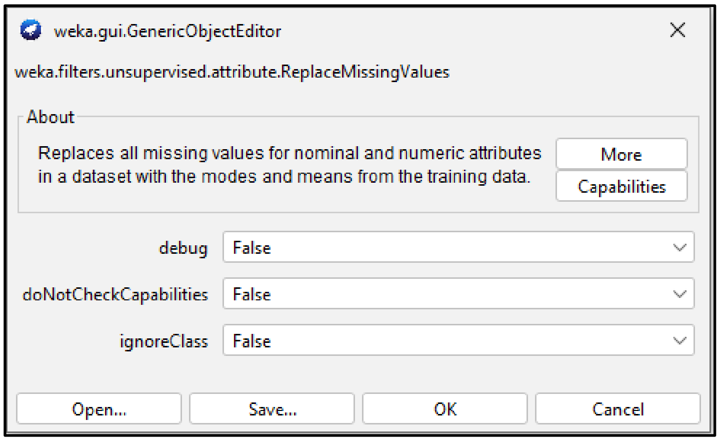
Before applying the “ReplaceMissingValues filter”



Relation: vote

No.	1: handicapped-infants Nominal	2: water-project-cost-sharing Nominal	3: adoption-of-the-budget-resolution Nominal	4: physician-fee-freeze Nominal	5: el-salvador-aid Nominal	6: religious-groups-in-schools Nominal	7: ant
1	n	y	n	y	y	y	n
2	n	y	n	y	y	y	n
3		y	y		y	y	n
4	n	y	y	n		y	
5	y	y	y	n	y	y	n
6	n	y	y	n	y	y	n
7	n	y	n	y	y	y	n
8	n	y	n	y	y	y	n
9	n	y	n	y	y	y	n
10	y	y	y	n	n	n	y
11	n	y	n	y	y	n	n
12	n	y	n	y	y	y	n
13	n	y	y	n	n	n	y
14	y	y	y	n	n	y	y
15	n	y	n	y	y	y	n
16	n	y	n	y	y	y	n
17	y	n	y	n	n	y	n
18	y		y	n	n	n	y
19	n	y	n	y	y	y	n
20	y	y	y	n	n	n	y
21	y	y	y	n	n		y
22	y	y	y	n	n	n	y
23	y		y	n	n	n	y

Right click (or left+



weka.gui.GenericObjectEditor

weka.filters.unsupervised.attribute.ReplaceMissingValues

About

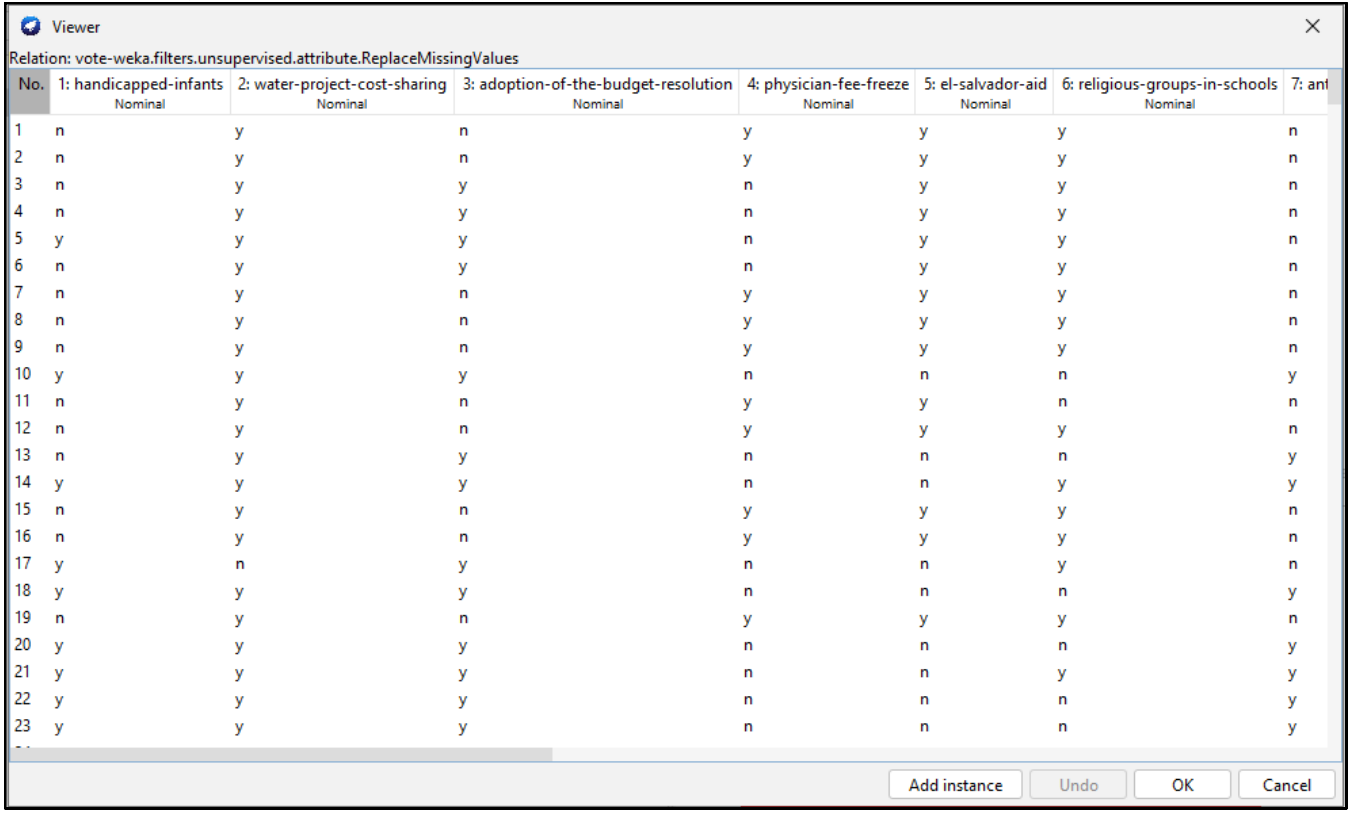
Replaces all missing values for nominal and numeric attributes in a dataset with the modes and means from the training data.

debug: False

doNotCheckCapabilities: False

ignoreClass: False

After applying the “ReplaceMissingValues filter”



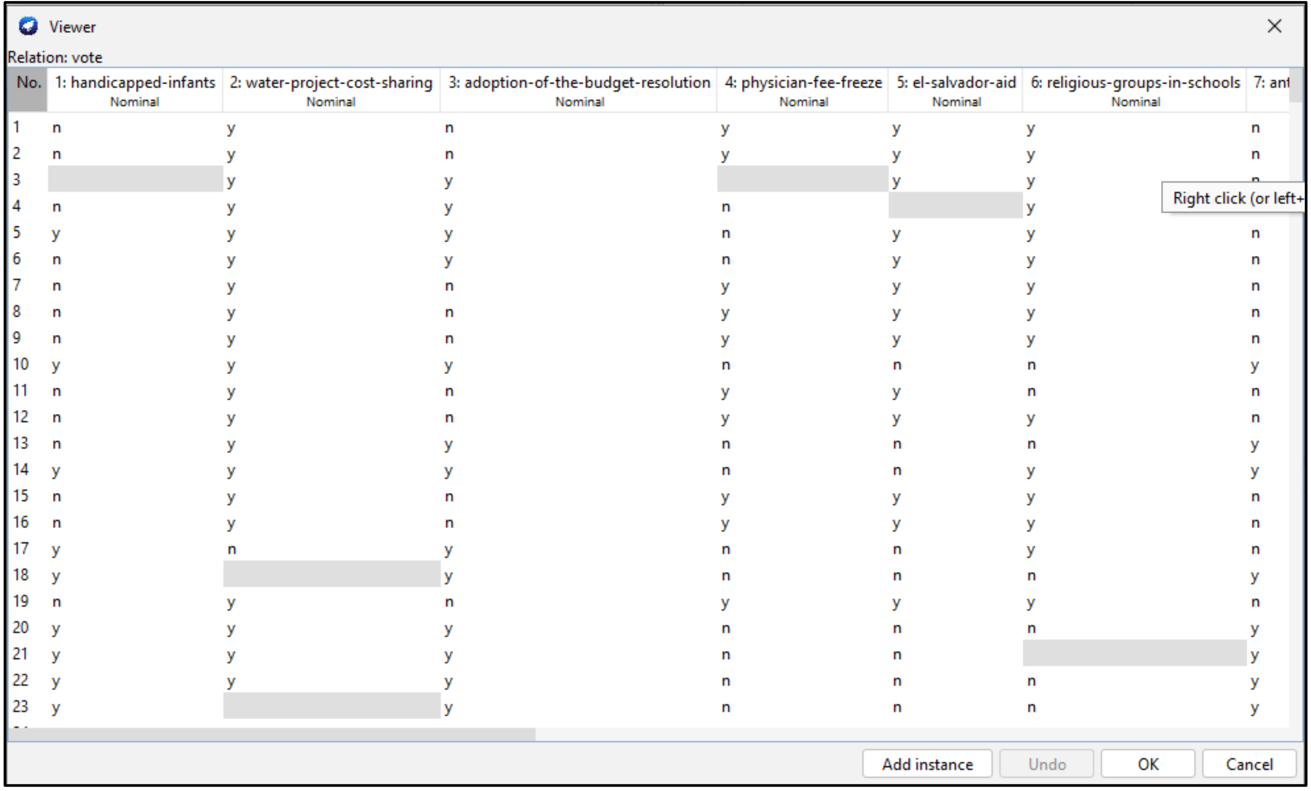
Viewer

Relation: vote-weka.filters.unsupervised.attribute.ReplaceMissingValues

No.	1: handicapped-infants Nominal	2: water-project-cost-sharing Nominal	3: adoption-of-the-budget-resolution Nominal	4: physician-fee-freeze Nominal	5: el-salvador-aid Nominal	6: religious-groups-in-schools Nominal	7: ant
1	n	y	n	y	y	y	n
2	n	y	n	y	y	y	n
3	n	y	y	n	y	y	n
4	n	y	y	n	y	y	n
5	y	y	y	n	y	y	n
6	n	y	y	n	y	y	n
7	n	y	n	y	y	y	n
8	n	y	n	y	y	y	n
9	n	y	n	y	y	y	n
10	y	y	y	n	n	n	y
11	n	y	n	y	y	n	n
12	n	y	n	y	y	y	n
13	n	y	y	n	n	n	y
14	y	y	y	n	n	y	y
15	n	y	n	y	y	y	n
16	n	y	n	y	y	y	n
17	y	n	y	n	n	y	n
18	y	y	y	n	n	n	y
19	n	y	n	y	y	y	n
20	y	y	y	n	n	n	y
21	y	y	y	n	n	y	y
22	y	y	y	n	n	n	y
23	y	y	y	n	n	n	y

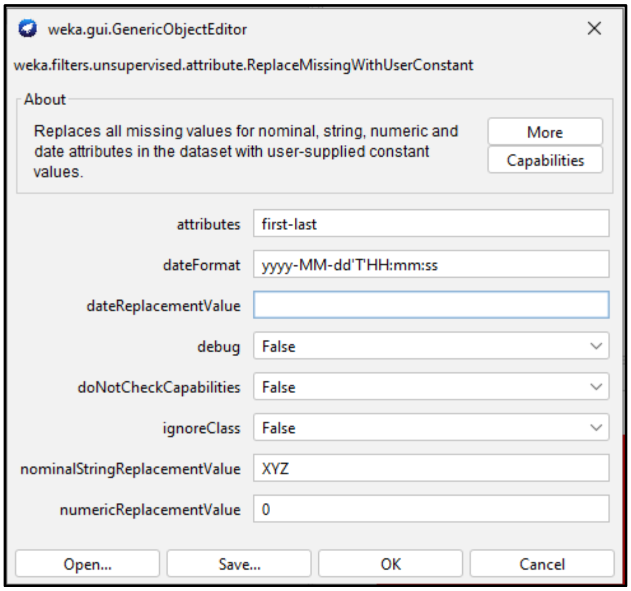
3. ReplaceMissingWithUserConstant filter: The Replace Missing With User Constant is an unsupervised attribute filter used to replace all missing values in a dataset with a constant value specified by the user. Allow users to define the custom constant to replace.

Before applying the "ReplaceMissingWithUserConstant filter"



Relation: vote

No.	1: handicapped-infants Nominal	2: water-project-cost-sharing Nominal	3: adoption-of-the-budget-resolution Nominal	4: physician-fee-freeze Nominal	5: el-salvador-aid Nominal	6: religious-groups-in-schools Nominal	7: ant
1	n	y	n	y	y	y	n
2	n	y	n	y	y	y	n
3		y	y		y	y	n
4	n	y	y	n		y	
5	y	y	y	n	y	y	n
6	n	y	y	n	y	y	n
7	n	y	n	y	y	y	n
8	n	y	n	y	y	y	n
9	n	y	n	y	y	y	n
10	y	y	y	n	n	n	y
11	n	y	n	y	y	n	n
12	n	y	n	y	y	y	n
13	n	y	y	n	n	n	y
14	y	y	y	n	n	y	y
15	n	y	n	y	y	y	n
16	n	y	n	y	y	y	n
17	y	n	y	n	n	y	n
18	y		y	n	n	n	y
19	n	y	n	y	y	y	n
20	y	y	y	n	n	n	y
21	y	y	y	n	n		y
22	y	y	y	n	n	n	y
23	y		y	n	n	n	y



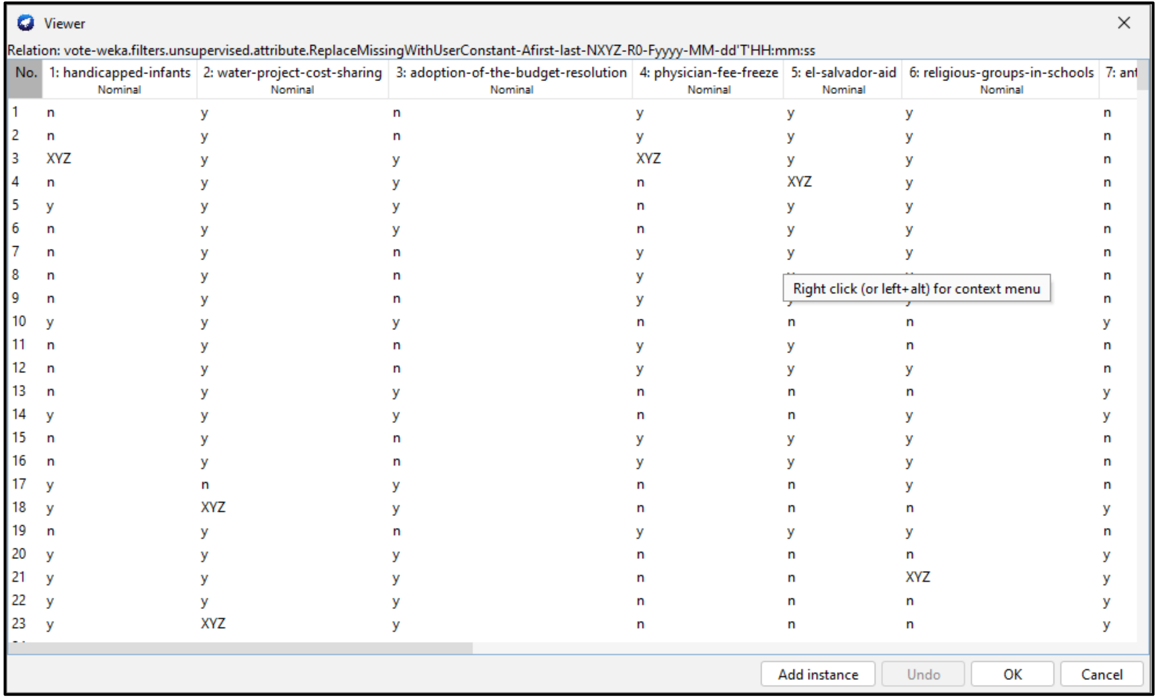
weka.filters.unsupervised.attribute.ReplaceMissingWithUserConstant

About

Replaces all missing values for nominal, string, numeric and date attributes in the dataset with user-supplied constant values.

attributes	first-last
dateFormat	yyyy-MM-dd'T'HH:mm:ss
dateReplacementValue	
debug	False
doNotCheckCapabilities	False
ignoreClass	False
nominalStringReplacementValue	XYZ
numericReplacementValue	0

After applying the “ReplaceMissingWithUserConstant filter”

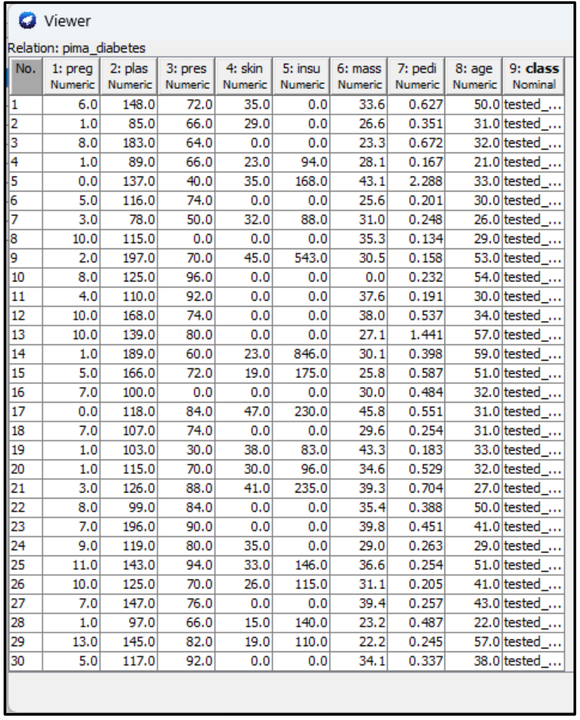


Relation: vote-weka.filters.unsupervised.attribute.ReplaceMissingWithUserConstant-Afirst-last-NXYZ-R0-Fyyyy-MM-dd'T'HH:mm:ss

No.	1: handicapped-infants Nominal	2: water-project-cost-sharing Nominal	3: adoption-of-the-budget-resolution Nominal	4: physician-fee-freeze Nominal	5: el-salvador-aid Nominal	6: religious-groups-in-schools Nominal	7: ant
1	n	y	n	y	y	y	n
2	n	y	n	y	y	y	n
3	XYZ	y	y	XYZ	y	y	n
4	n	y	y	n	XYZ	y	n
5	y	y	y	n	y	y	n
6	n	y	y	n	y	y	n
7	n	y	n	y	y	y	n
8	n	y	n	y			n
9	n	y	n	y			n
10	y	y	y	n	n	n	y
11	n	y	n	y	y	n	n
12	n	y	n	y	y	y	n
13	n	y	y	n	n	n	y
14	y	y	y	n	n	y	y
15	n	y	n	y	y	y	n
16	n	y	n	y	y	y	n
17	y	n	y	n	n	y	n
18	y	XYZ	y	n	n	n	y
19	n	y	n	y	y	y	n
20	y	y	y	n	n	n	y
21	y	y	y	n	n	XYZ	y
22	y	y	y	n	n	n	y
23	y	XYZ	y	n	n	n	y

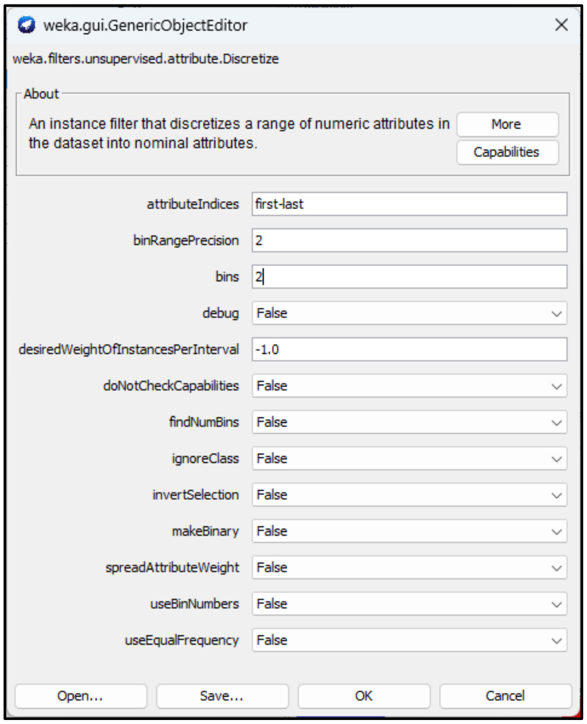
4. Descritize filter: The Descritize filter is an unsupervised attribute into nominal attributes by dividing the numeric range into discrete intervals or bins. Descritization helps simplify data and can improve model performance in certain scenarios.

Before applying the “Descritize filter”



Relation: pima_diabetes

No.	1: preg Numeric	2: plas Numeric	3: pres Numeric	4: skin Numeric	5: insu Numeric	6: mass Numeric	7: pedi Numeric	8: age Numeric	9: **class** Nominal
1	6.0	148.0	72.0	35.0	0.0	33.6	0.627	50.0	tested_...
2	1.0	85.0	66.0	29.0	0.0	26.6	0.351	31.0	tested_...
3	8.0	183.0	64.0	0.0	0.0	23.3	0.672	32.0	tested_...
4	1.0	89.0	66.0	23.0	94.0	28.1	0.167	21.0	tested_...
5	0.0	137.0	40.0	35.0	168.0	43.1	2.288	33.0	tested_...
6	5.0	116.0	74.0	0.0	0.0	25.6	0.201	30.0	tested_...
7	3.0	78.0	50.0	32.0	88.0	31.0	0.248	26.0	tested_...
8	10.0	115.0	0.0	0.0	0.0	35.3	0.134	29.0	tested_...
9	2.0	197.0	70.0	45.0	543.0	30.5	0.158	53.0	tested_...
10	8.0	125.0	96.0	0.0	0.0	0.0	0.232	54.0	tested_...
11	4.0	110.0	92.0	0.0	0.0	37.6	0.191	30.0	tested_...
12	10.0	168.0	74.0	0.0	0.0	38.0	0.537	34.0	tested_...
13	10.0	139.0	80.0	0.0	0.0	27.1	1.441	57.0	tested_...
14	1.0	189.0	60.0	23.0	846.0	30.1	0.398	59.0	tested_...
15	5.0	166.0	72.0	19.0	175.0	25.8	0.587	51.0	tested_...
16	7.0	100.0	0.0	0.0	0.0	30.0	0.484	32.0	tested_...
17	0.0	118.0	84.0	47.0	230.0	45.8	0.551	31.0	tested_...
18	7.0	107.0	74.0	0.0	0.0	29.6	0.254	31.0	tested_...
19	1.0	103.0	30.0	38.0	83.0	43.3	0.183	33.0	tested_...
20	1.0	115.0	70.0	30.0	96.0	34.6	0.529	32.0	tested_...
21	3.0	126.0	88.0	41.0	235.0	39.3	0.704	27.0	tested_...
22	8.0	99.0	84.0	0.0	0.0	35.4	0.388	50.0	tested_...
23	7.0	196.0	90.0	0.0	0.0	39.8	0.451	41.0	tested_...
24	9.0	119.0	80.0	35.0	0.0	29.0	0.263	29.0	tested_...
25	11.0	143.0	94.0	33.0	146.0	36.6	0.254	51.0	tested_...
26	10.0	125.0	70.0	26.0	115.0	31.1	0.205	41.0	tested_...
27	7.0	147.0	76.0	0.0	0.0	39.4	0.257	43.0	tested_...
28	1.0	97.0	66.0	15.0	140.0	23.2	0.487	22.0	tested_...
29	13.0	145.0	82.0	19.0	110.0	22.2	0.245	57.0	tested_...
30	5.0	117.0	92.0	0.0	0.0	34.1	0.337	38.0	tested_...



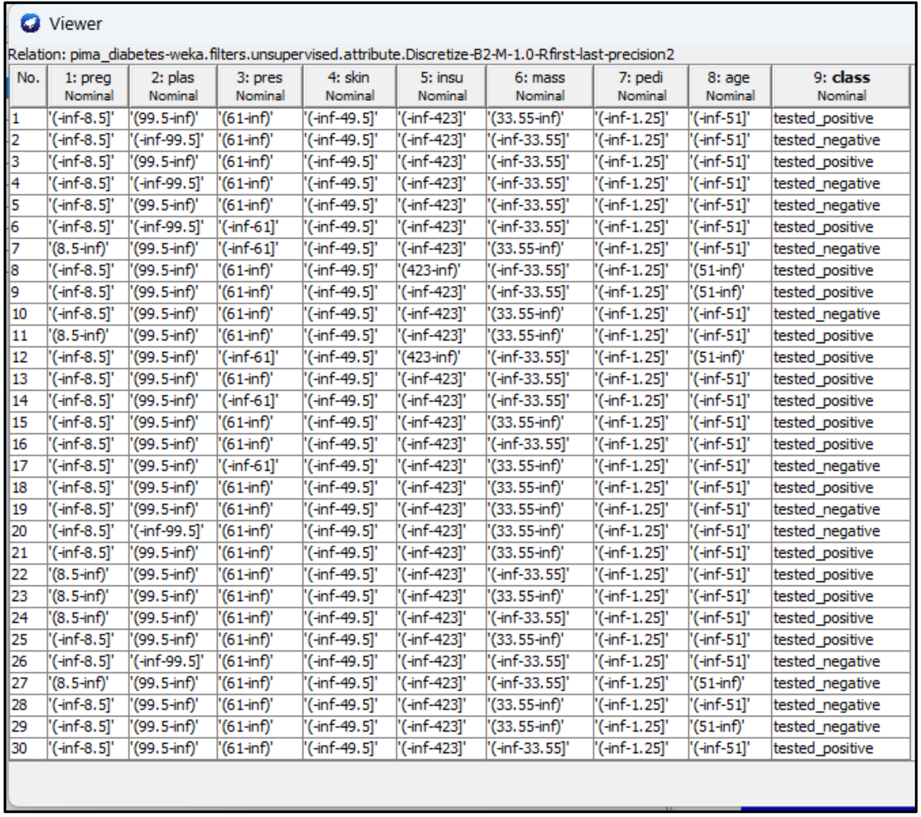
weka.filters.unsupervised.attribute.Discretize

About

An instance filter that discretizes a range of numeric attributes in the dataset into nominal attributes.

attributeIndices	first-last
binRangePrecision	2
bins	2
debug	False
desiredWeightOfInstancesPerInterval	-1.0
doNotCheckCapabilities	False
findNumBins	False
ignoreClass	False
invertSelection	False
makeBinary	False
spreadAttributeWeight	False
useBinNumbers	False
useEqualFrequency	False

After applying the “Descritize filter”



Relation: pima_diabetes-weka.filters.unsupervised.attribute.Discretize-B2-M-1.0-Rfirst-last-precision2

No.	1: preg	2: plas	3: pres	4: skin	5: insu	6: mass	7: pedi	8: age	9: **class**
	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal
1	'(-inf-8.5]'	'(99.5-inf)'	'(61-inf)'	'(-inf-49.5]'	'(-inf-423]'	'(33.55-inf)'	'(-inf-1.25]'	'(-inf-51]'	tested_positive
2	'(-inf-8.5]'	'(-inf-99.5]'	'(61-inf)'	'(-inf-49.5]'	'(-inf-423]'	'(-inf-33.55]'	'(-inf-1.25]'	'(-inf-51]'	tested_negative
3	'(-inf-8.5]'	'(99.5-inf)'	'(61-inf)'	'(-inf-49.5]'	'(-inf-423]'	'(-inf-33.55]'	'(-inf-1.25]'	'(-inf-51]'	tested_positive
4	'(-inf-8.5]'	'(-inf-99.5]'	'(61-inf)'	'(-inf-49.5]'	'(-inf-423]'	'(-inf-33.55]'	'(-inf-1.25]'	'(-inf-51]'	tested_negative
5	'(-inf-8.5]'	'(99.5-inf)'	'(61-inf)'	'(-inf-49.5]'	'(-inf-423]'	'(-inf-33.55]'	'(-inf-1.25]'	'(-inf-51]'	tested_negative
6	'(-inf-8.5]'	'(-inf-99.5]'	'(-inf-61]'	'(-inf-49.5]'	'(-inf-423]'	'(-inf-33.55]'	'(-inf-1.25]'	'(-inf-51]'	tested_positive
7	'(8.5-inf)'	'(99.5-inf)'	'(-inf-61]'	'(-inf-49.5]'	'(-inf-423]'	'(33.55-inf)'	'(-inf-1.25]'	'(-inf-51]'	tested_negative
8	'(-inf-8.5]'	'(99.5-inf)'	'(61-inf)'	'(-inf-49.5]'	'(423-inf)'	'(-inf-33.55]'	'(-inf-1.25]'	'(51-inf)'	tested_positive
9	'(-inf-8.5]'	'(99.5-inf)'	'(61-inf)'	'(-inf-49.5]'	'(-inf-423]'	'(-inf-33.55]'	'(-inf-1.25]'	'(51-inf)'	tested_positive
10	'(-inf-8.5]'	'(99.5-inf)'	'(61-inf)'	'(-inf-49.5]'	'(-inf-423]'	'(33.55-inf)'	'(-inf-1.25]'	'(-inf-51]'	tested_negative
11	'(8.5-inf)'	'(99.5-inf)'	'(61-inf)'	'(-inf-49.5]'	'(-inf-423]'	'(33.55-inf)'	'(-inf-1.25]'	'(-inf-51]'	tested_positive
12	'(-inf-8.5]'	'(99.5-inf)'	'(-inf-61]'	'(-inf-49.5]'	'(423-inf)'	'(-inf-33.55]'	'(-inf-1.25]'	'(51-inf)'	tested_positive
13	'(-inf-8.5]'	'(99.5-inf)'	'(61-inf)'	'(-inf-49.5]'	'(-inf-423]'	'(-inf-33.55]'	'(-inf-1.25]'	'(-inf-51]'	tested_positive
14	'(-inf-8.5]'	'(99.5-inf)'	'(-inf-61]'	'(-inf-49.5]'	'(-inf-423]'	'(-inf-33.55]'	'(-inf-1.25]'	'(-inf-51]'	tested_positive
15	'(-inf-8.5]'	'(99.5-inf)'	'(61-inf)'	'(-inf-49.5]'	'(-inf-423]'	'(33.55-inf)'	'(-inf-1.25]'	'(-inf-51]'	tested_positive
16	'(-inf-8.5]'	'(99.5-inf)'	'(61-inf)'	'(-inf-49.5]'	'(-inf-423]'	'(-inf-33.55]'	'(-inf-1.25]'	'(-inf-51]'	tested_positive
17	'(-inf-8.5]'	'(99.5-inf)'	'(-inf-61]'	'(-inf-49.5]'	'(-inf-423]'	'(33.55-inf)'	'(-inf-1.25]'	'(-inf-51]'	tested_negative
18	'(-inf-8.5]'	'(99.5-inf)'	'(61-inf)'	'(-inf-49.5]'	'(-inf-423]'	'(33.55-inf)'	'(-inf-1.25]'	'(-inf-51]'	tested_positive
19	'(-inf-8.5]'	'(99.5-inf)'	'(61-inf)'	'(-inf-49.5]'	'(-inf-423]'	'(33.55-inf)'	'(-inf-1.25]'	'(-inf-51]'	tested_negative
20	'(-inf-8.5]'	'(-inf-99.5]'	'(61-inf)'	'(-inf-49.5]'	'(-inf-423]'	'(33.55-inf)'	'(-inf-1.25]'	'(-inf-51]'	tested_negative
21	'(-inf-8.5]'	'(99.5-inf)'	'(61-inf)'	'(-inf-49.5]'	'(-inf-423]'	'(33.55-inf)'	'(-inf-1.25]'	'(-inf-51]'	tested_positive
22	'(8.5-inf)'	'(99.5-inf)'	'(61-inf)'	'(-inf-49.5]'	'(-inf-423]'	'(-inf-33.55]'	'(-inf-1.25]'	'(-inf-51]'	tested_positive
23	'(8.5-inf)'	'(99.5-inf)'	'(61-inf)'	'(-inf-49.5]'	'(-inf-423]'	'(33.55-inf)'	'(-inf-1.25]'	'(-inf-51]'	tested_positive
24	'(8.5-inf)'	'(99.5-inf)'	'(61-inf)'	'(-inf-49.5]'	'(-inf-423]'	'(-inf-33.55]'	'(-inf-1.25]'	'(-inf-51]'	tested_positive
25	'(-inf-8.5]'	'(99.5-inf)'	'(61-inf)'	'(-inf-49.5]'	'(-inf-423]'	'(33.55-inf)'	'(-inf-1.25]'	'(-inf-51]'	tested_positive
26	'(-inf-8.5]'	'(-inf-99.5]'	'(61-inf)'	'(-inf-49.5]'	'(-inf-423]'	'(-inf-33.55]'	'(-inf-1.25]'	'(-inf-51]'	tested_negative
27	'(8.5-inf)'	'(99.5-inf)'	'(61-inf)'	'(-inf-49.5]'	'(-inf-423]'	'(-inf-33.55]'	'(-inf-1.25]'	'(51-inf)'	tested_negative
28	'(-inf-8.5]'	'(99.5-inf)'	'(61-inf)'	'(-inf-49.5]'	'(-inf-423]'	'(33.55-inf)'	'(-inf-1.25]'	'(-inf-51]'	tested_negative
29	'(-inf-8.5]'	'(99.5-inf)'	'(61-inf)'	'(-inf-49.5]'	'(-inf-423]'	'(33.55-inf)'	'(-inf-1.25]'	'(51-inf)'	tested_negative
30	'(-inf-8.5]'	'(99.5-inf)'	'(61-inf)'	'(-inf-49.5]'	'(-inf-423]'	'(-inf-33.55]'	'(-inf-1.25]'	'(-inf-51]'	tested_positive