

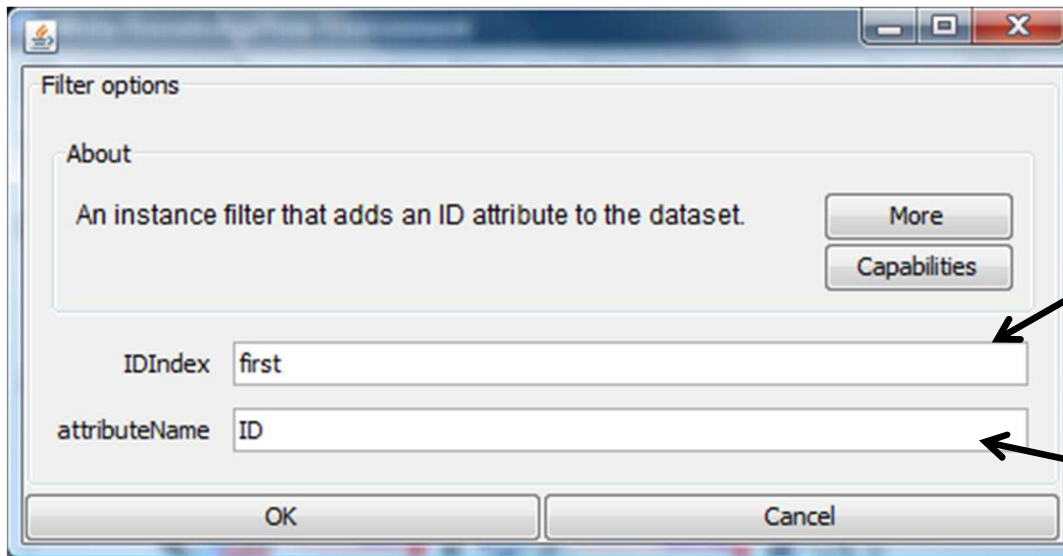


Wstępne przetwarzanie danych

Weka

AddID

- Dodaje nowy atrybut, który jest unikalnym identyfikatorem wiersza.

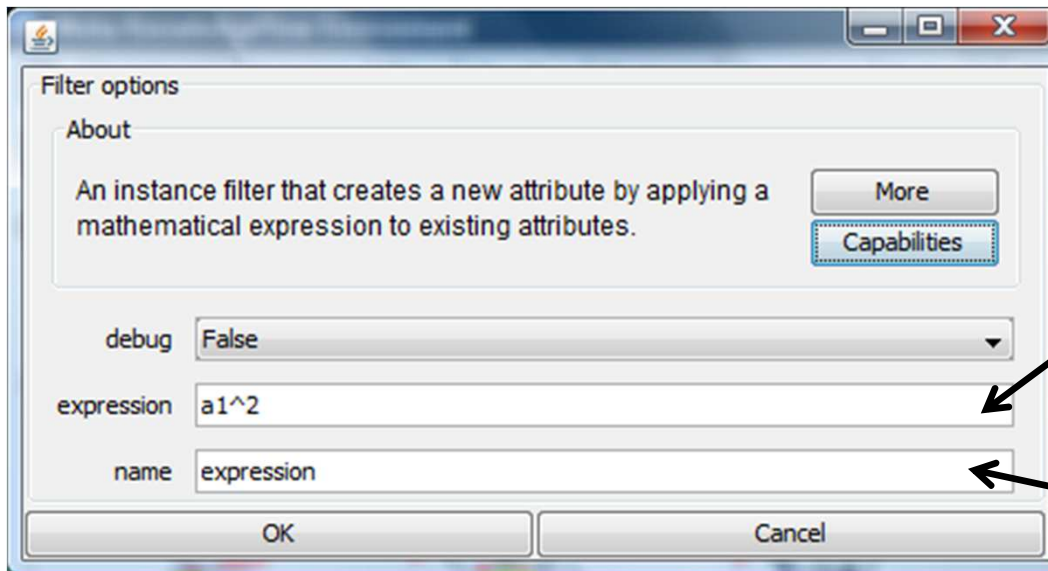
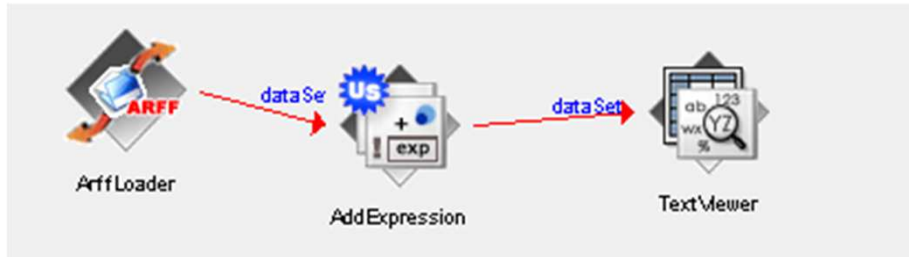


Pozycja na której wstawiany jest nowy atrybut

Nazwa nowego atrybutu

AddExpression

- Dodaje nowy atrybut wyliczany na podstawie pozostałych.

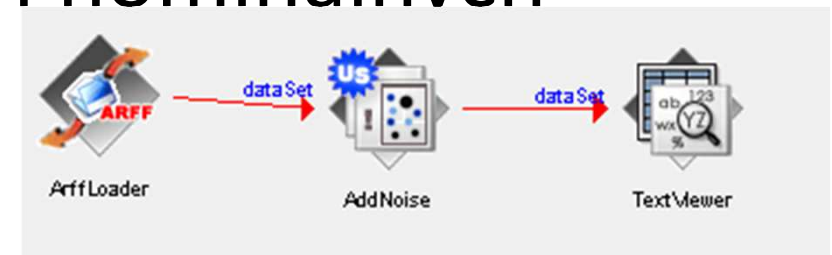
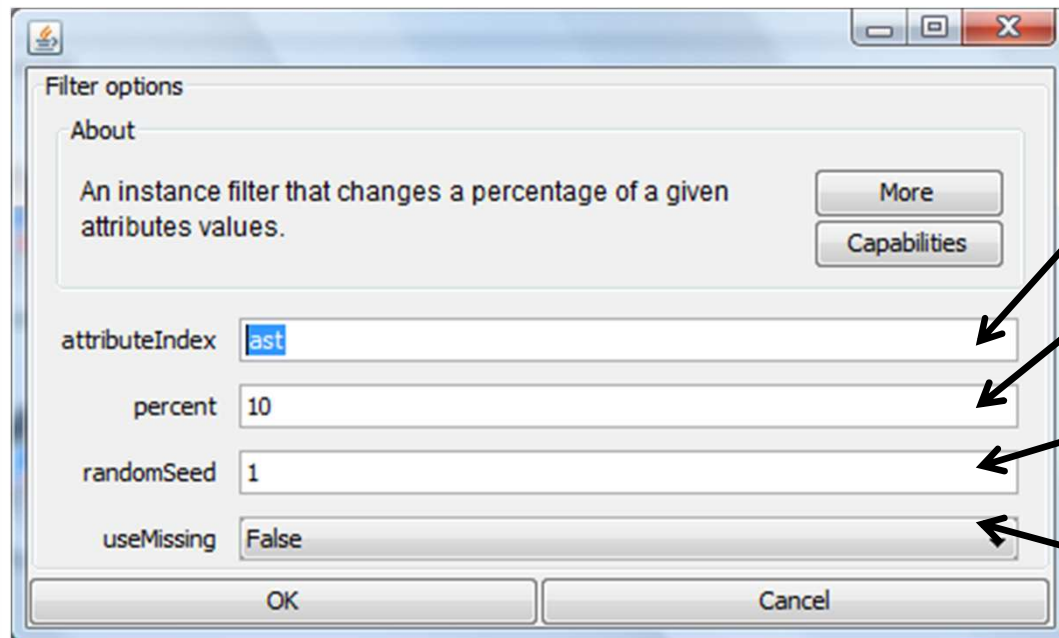


Wyrażenie wyliczające
nowy atrybut

Nazwa nowego atrybutu

AddNoise

- Wybiera losowo zadany procent wierszy.
- W wybranych wierszach zamienia wartość wybranego atrybutu na inną.
- Działa tylko na atrybutach nominalnych



Numer modyfikowanego atrybutu

Jaki procent wierszy zmienić?

Ziarno generatora liczb pseudolowowych

Czy wstawiać wartości puste?

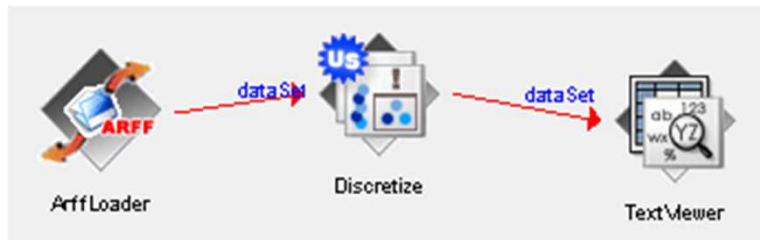
Discretize (1) - equalWidth

- 5, 7, 12, 35, 65, 82, 84, 88, 90, 95
- $(95-5)/5=18$
 - $(-\infty, 23]$
 - $(23, 41]$
 - $(41, 59]$
 - $(59, 77]$
 - $(77, +\infty)$
- 5, 7, 12, 35, 65, 82, 84, 88, 90, 95

Discretize (2) - equalFrequency

- 5, 7, 12, 35, 65, 82, 84, 88, 90, 95
- $10/5=2$
 - [-inf,9.5)
 - [9.5,50)
 - [50,83)
 - [83,89)
 - [89,inf+)
- 5, 7, 12, 35, 65, 82, 84, 88, 90, 95

Discretize (3)



The screenshot shows the 'Discretize' dialog box with the following settings:

- Filter options**
 - About**
 - An instance filter that discretizes a range of numeric attributes in the dataset into nominal attributes.
 - More**
 - Capabilities**
- attributeIndices**: first-last
- bins**: 10
- desiredWeightOfInstancesPerInterval**: -1.0
- findNumBins**: False
- ignoreClass**: False
- invertSelection**: False
- makeBinary**: False
- useEqualFrequency**: True

At the bottom are **OK** and **Cancel** buttons.

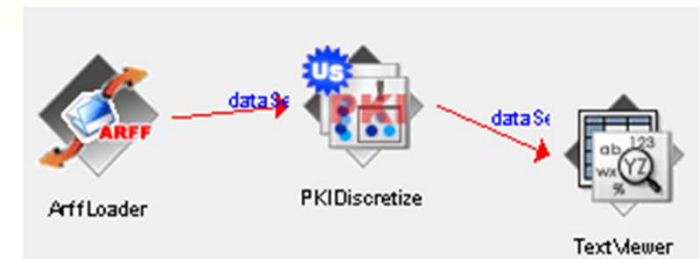
Które atrybuty przerabiamy

Ile binów

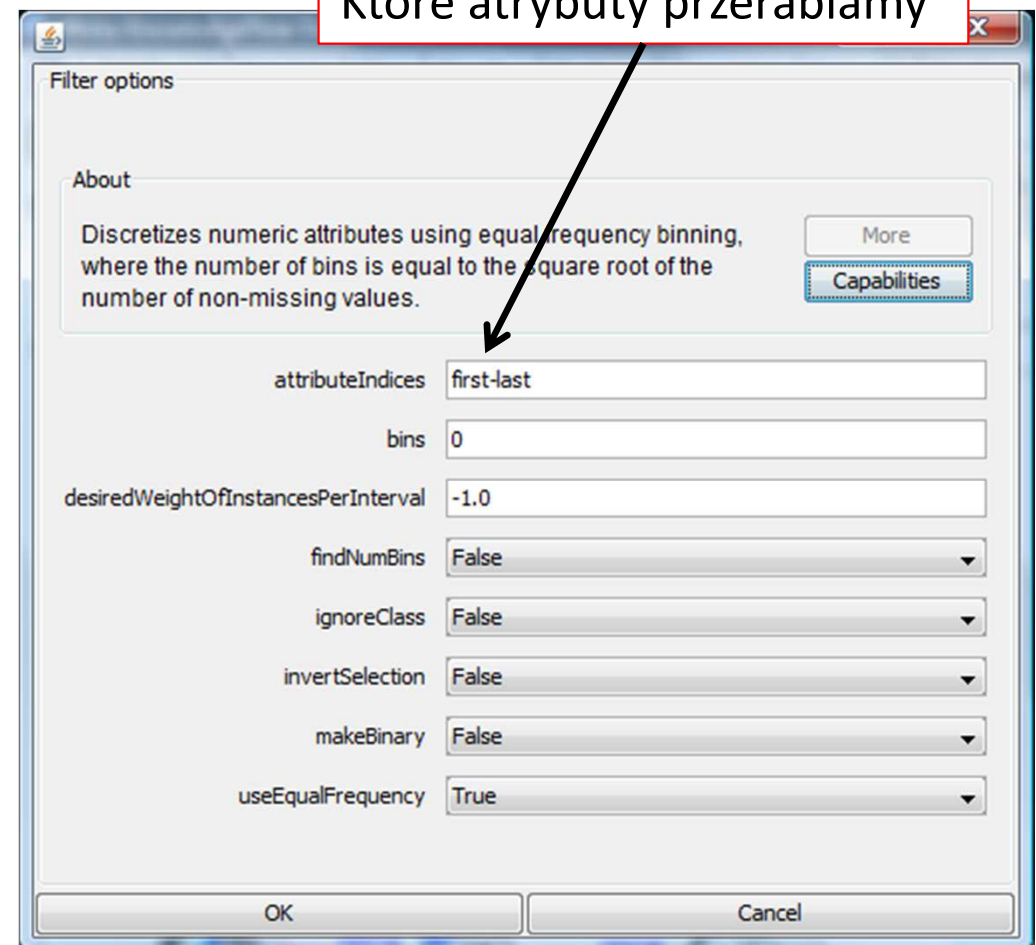
Używać equalFrequency
(domyślnie equalWidth)

PKIDiscretize

- Equal Frequency Binning
- Oblicza liczbę binów jako pierwiastek z liczby różnych (wykorzystanych) wartości w atrybucie.



Które atrybuty przerabiamy

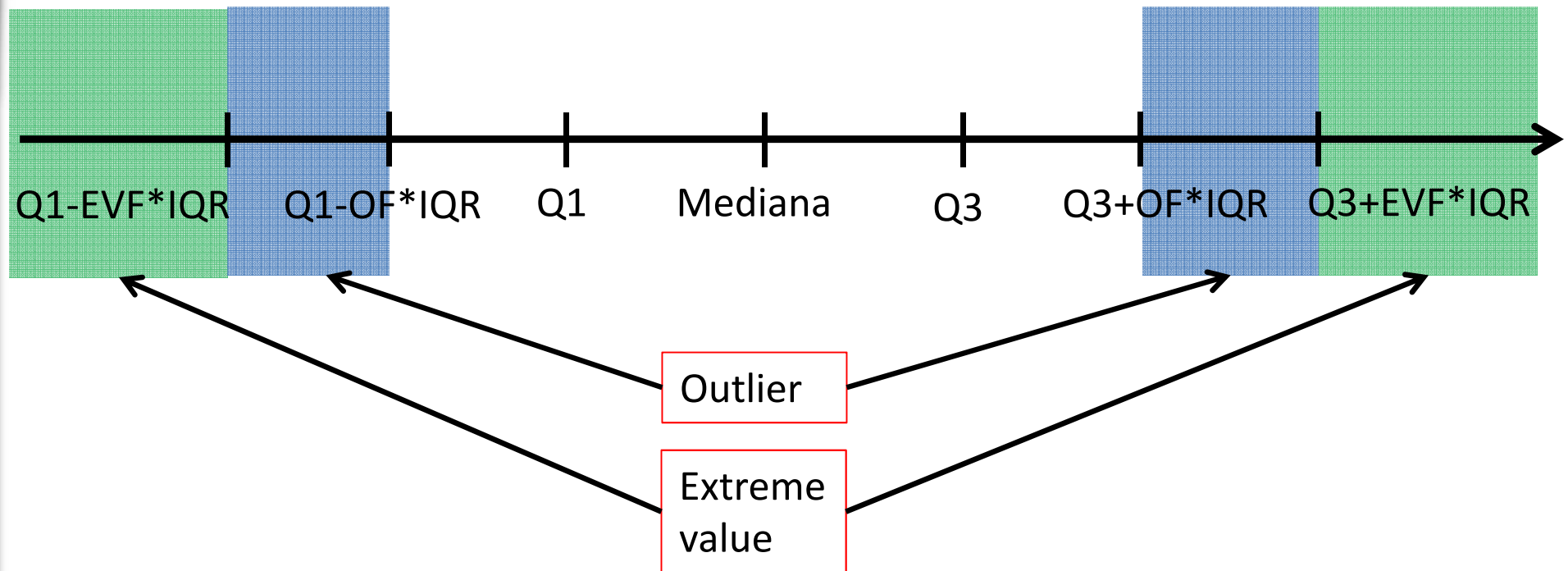


InterQuartileRange (1)

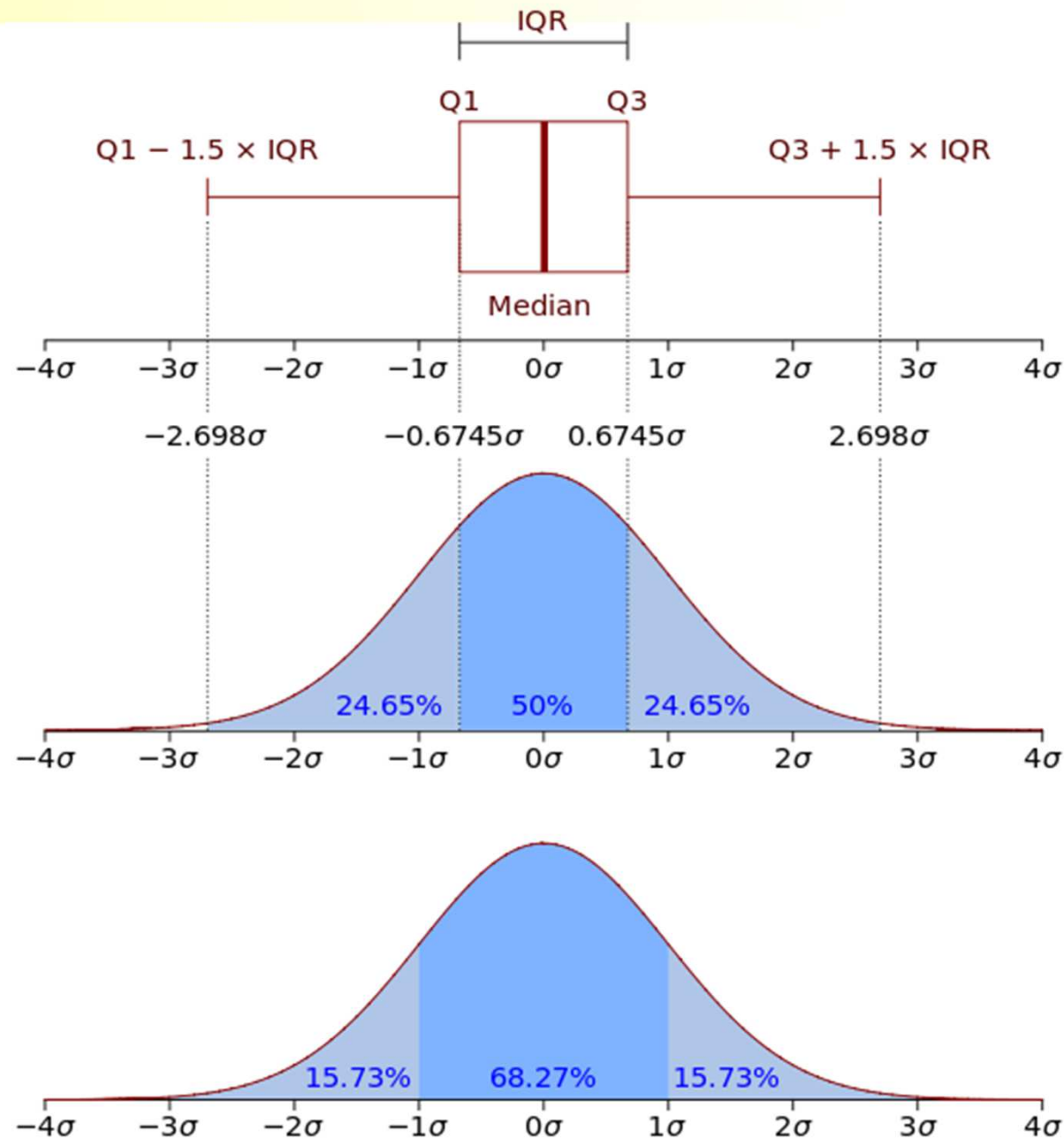
- Kwartył
 - dolny $x_{1/4}$ – próbek mniejszych niż $x_{1/4}$ jest 25% (inne oznaczenie Q1)
 - mediana $x_{1/2}$ – próbek mniejszych niż $x_{1/2}$ jest 50%
 - górny $x_{3/4}$ – próbek mniejszych niż $x_{3/4}$ jest 75% (inne oznaczenie Q3)
- Rozstęp kwartylny (IQR) = Q3-Q1

InterQuartileRange (2)

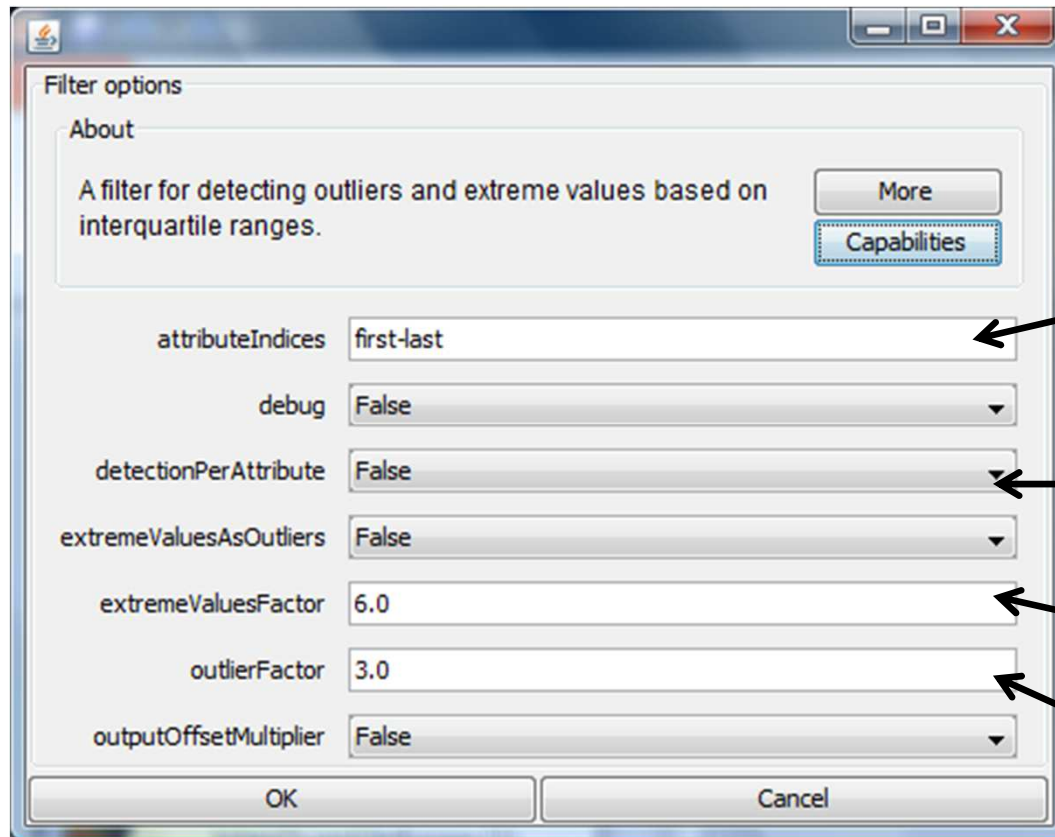
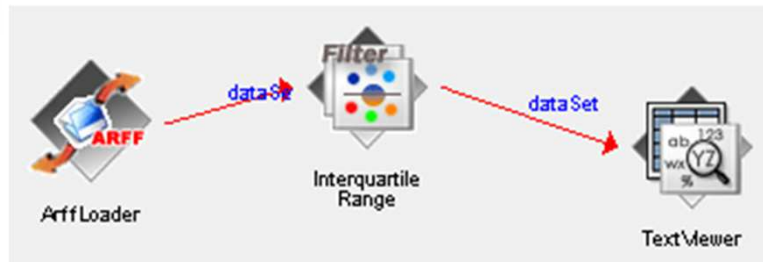
- Outlier
 - Outlier factor (OF) np. 3
 - Extreme value factor (EVF) np. 6



InterQuartileRange (3)



InterQuartileRange (4)



Które atrybuty bierzemy pod uwagę

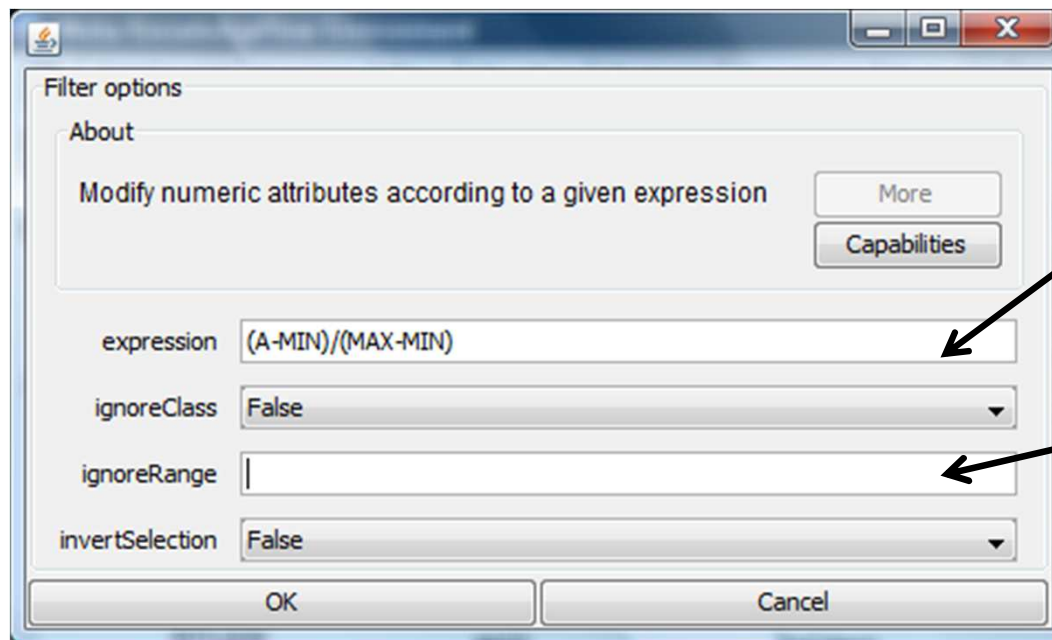
Sprawdzamy dla każdego atrybutu numerycznego

EVF

OF

MathExpression

- Przekształca atrybut poprzez podstawienie do wzoru



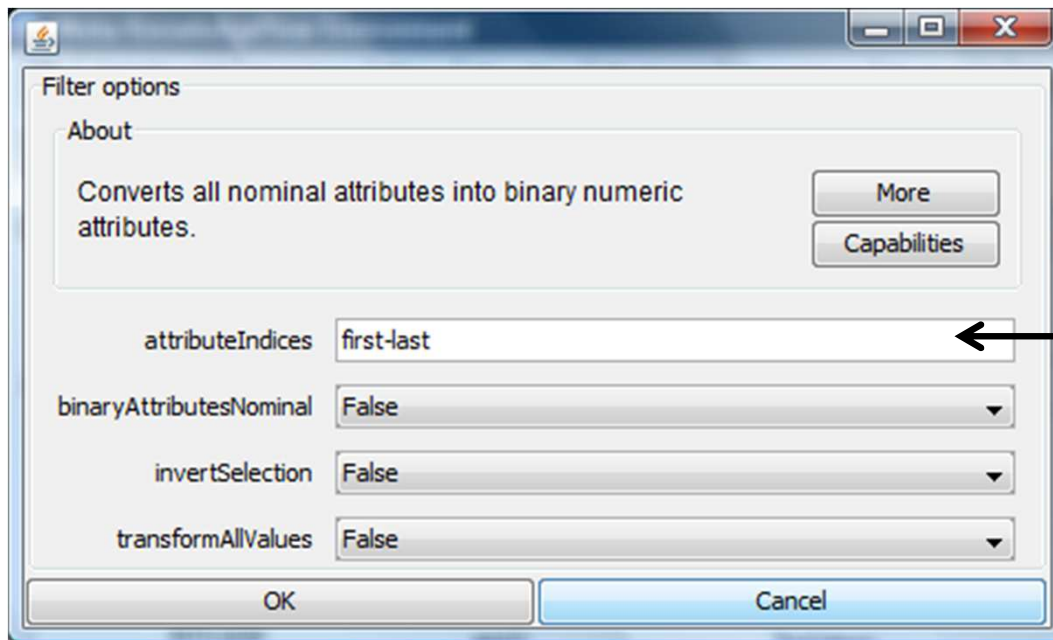
Wzór

Których atrybutów **nie** bierzemy pod uwagę

NominalToBinary (1)

- Atrybut nominalny: wartości ze skończonego zbioru bez relacji porządku, np.: KOLORY: (czerwony, zielony, niebieski, żółty)
- Atrybut binarny: dwie wartości 0 i 1.
- Konwersja: tyle atrybutów binarnych ile wartości atrybutu nominalnego
 - Atrybut CzyCzerwony? 0/1
 - Atrybut CzyZielony? 0/1
 - Atrybut CzyNiebieski? 0/1
 - Atrybut CzyŻółty? 0/1

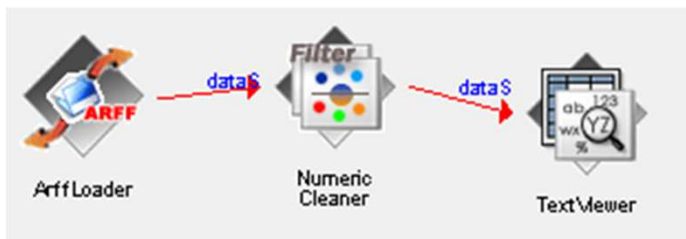
NominalToBinary (2)



Które atrybuty

NumericCleaner

- Zamienia liczby bliskie x na y
- Zamienia liczby większe / mniejsze od progu na zadaną wartość.



The screenshot shows the 'NumericCleaner' dialog box with the following fields and annotations:

- Filter options** (Section Header)
- About** (Section Header)
- A filter that 'cleanses' the numeric data from values that are too small, too big or very close to a certain value (e.** (Text)
- More** (Button)
- Capabilities** (Button)
- attributeIndices** (Text field): first-last (Annotated with 'Które atrybuty?')
- closeTo** (Text field): 0.0 (Annotated with 'x')
- closeToDefault** (Text field): 0.0 (Annotated with 'y')
- closeToTolerance** (Text field): 1.0E-6 (Annotated with 'Co to znaczy bliskie?')
- debug** (Text field): False
- decimals** (Text field): -1
- includeClass** (Text field): False
- invertSelection** (Text field): False
- maxDefault** (Text field): 1.7976931348623157E308
- maxThreshold** (Text field): 1.7976931348623157E308 (Annotated with 'Progi/wartości Min/Max')
- minDefault** (Text field): -1.7976931348623157E308
- minThreshold** (Text field): -1.7976931348623157E308
- OK** (Button)
- Cancel** (Button)

Obfuscate

- Zmienia nazwy atrybutów i wartości nominalnych.
- Brak parametrów konfiguracyjnych



Standardize

- Standaryzuje wszystkie atrybuty numeryczne.

$$z = \frac{x - \mu}{\sigma}$$



- Atrybut po standaryzacji ma średnią równą 0 i odchylenie standardowe równe 1.

RemoveFrequentValues

- Usuwa wiersze, w których atrybut posiada wartości rzadkie/częste.

The image displays a workflow in the Weka GUI and a detailed view of the 'RemoveFrequentValues' filter options.

Workflow: The workflow starts with an 'ArffLoader' node, followed by an 'AddID' node, then a 'Filter' node (labeled 'Remove Frequent Values'), and finally a 'Text Viewer' node. Arrows labeled 'dataSet' connect these nodes. A 'Text Viewer' node is also shown above the 'AddID' node, connected by a red arrow.

Filter options dialog: The dialog is titled 'Filter options' and contains the following settings:

- About:** Determines which values (frequent or infrequent ones) of an (nominal) attribute are retained and filters the instances accordingly.
- attributeIndex:** last
- invertSelection:** False
- modifyHeader:** False
- numValues:** 2
- useLeastValues:** False

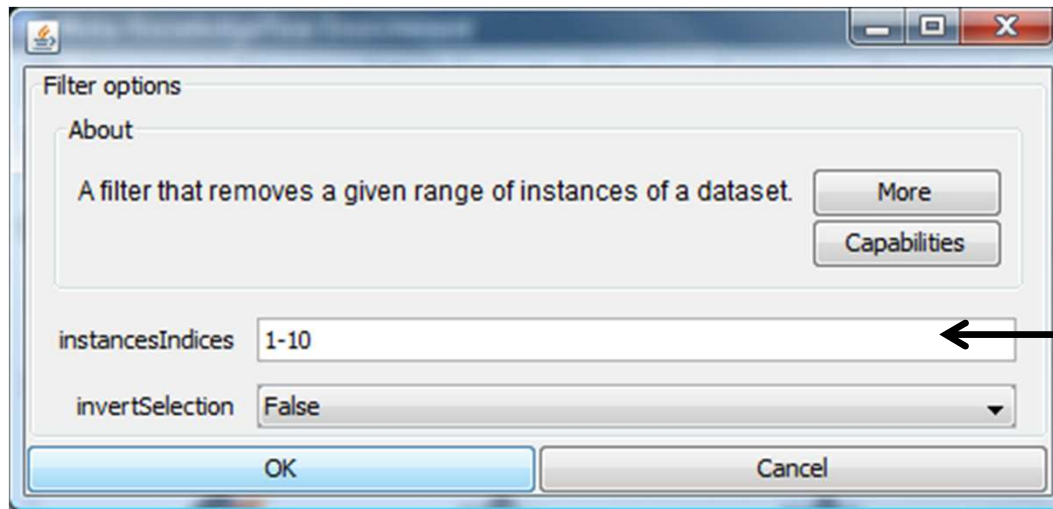
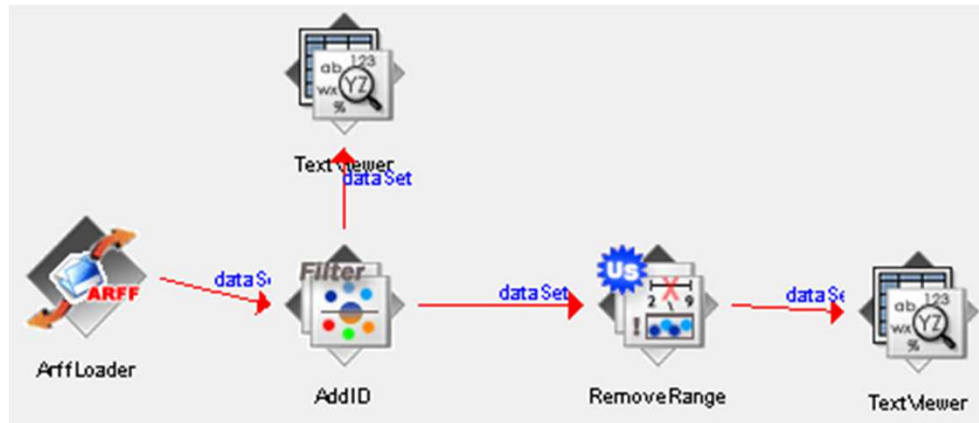
Buttons for 'More' and 'Capabilities' are visible on the right side of the dialog.

Annotations:

- Które atrybut analizujemy?
- Ile najczęstszych/najrzadszych zostawiamy?
- Zostawiamy najczęstsze czy najrzadsze?

RemoveRange

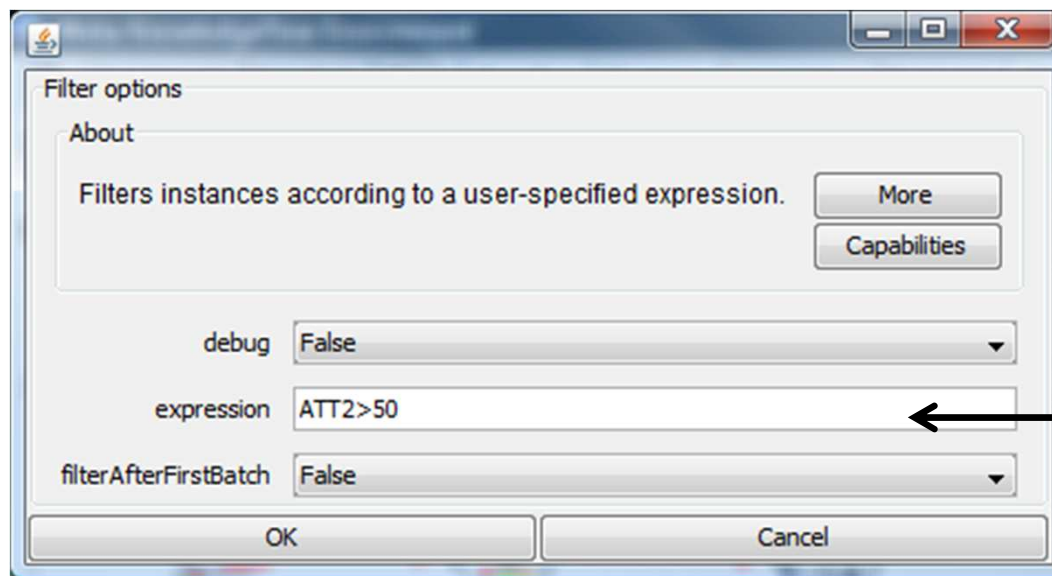
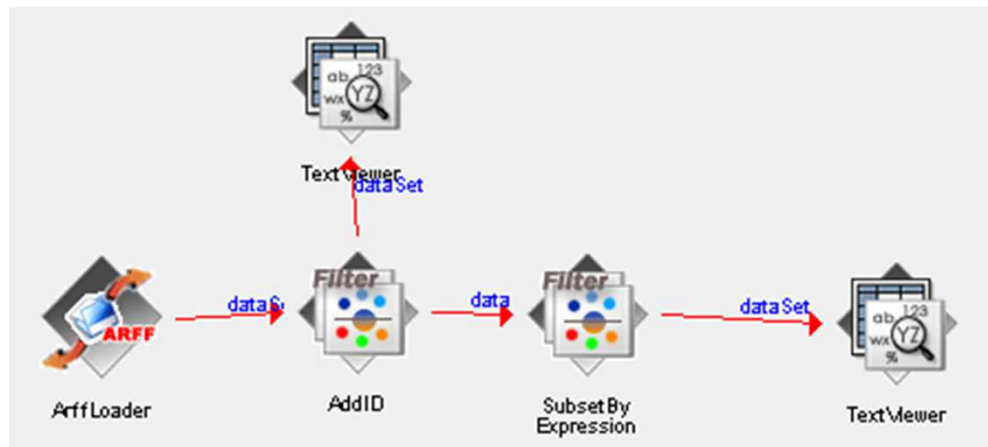
- Usuwa wybrane rekordy



Które rekordy

SubsetByExpression

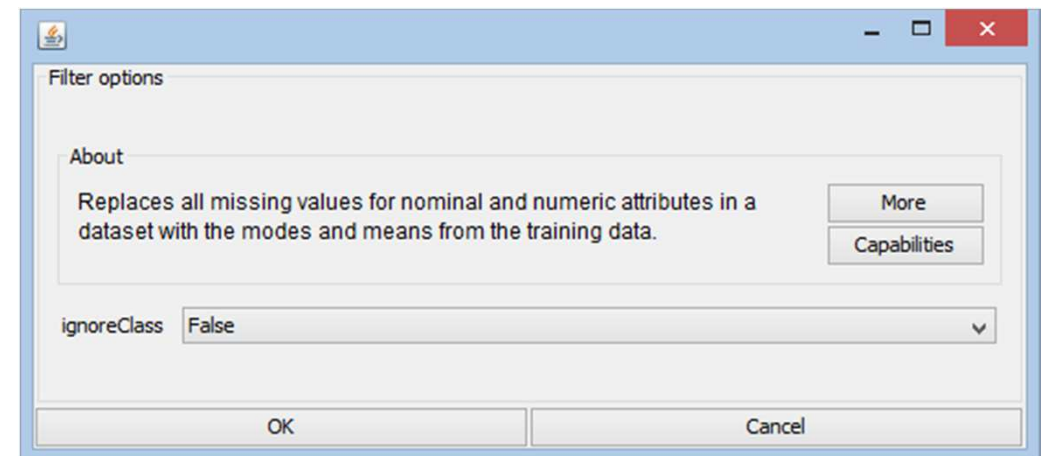
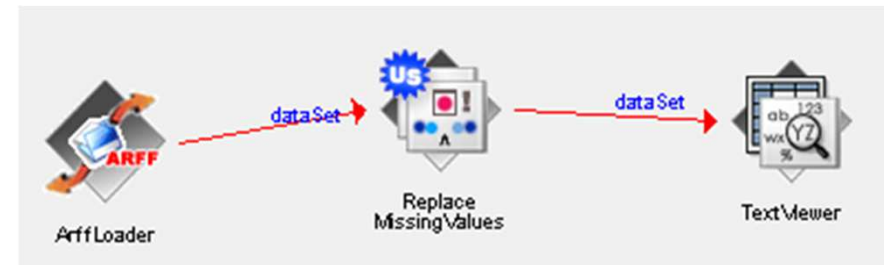
- Usuwa rekordy nie spełniające warunku



Warunek

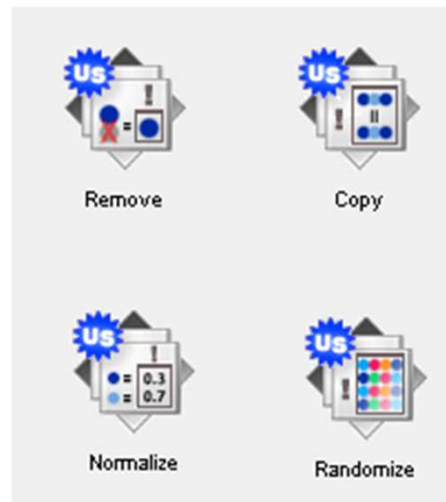
ReplaceMissingValues

- Uzupełnia nieznane wartości
- Wartość liczbowa wyliczana jako średnia
- Wartość nominalna wyliczana jako wartość najczęściej występująca



Zadania

- Rozpracuj jak działają następujące filtry w wece:



- Używając ReplaceMissingValues, InterQuantileRange i SubsetByExpression znajdź nietypowe przypadki w bazie