

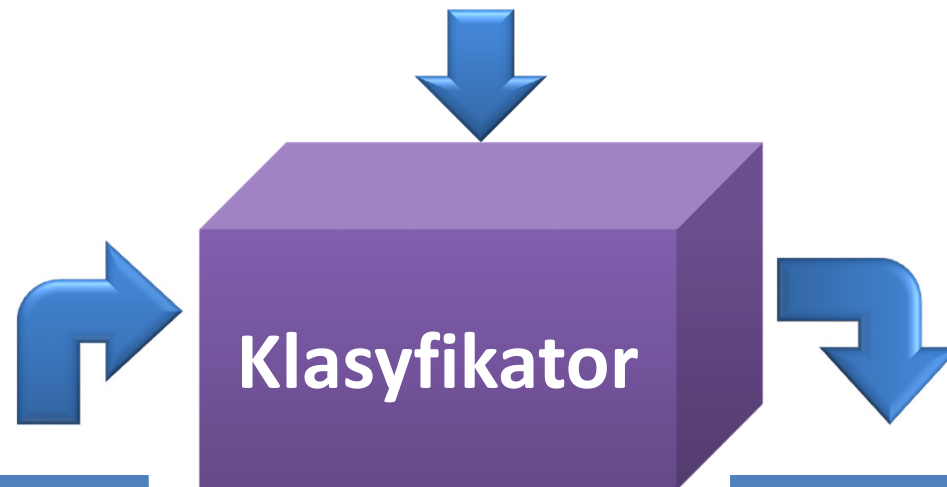


Określanie ważności atrybutów

Rapid Miner

Klasyfikacja (1/2)

TEMP	BÓL	WYSYPKA	GARDŁO	DIAGNOZA
36.6	T	BRAK	NORMA	NIESTRAWNOŚĆ
37.5	N	MAŁA	PRZEKR.	ALERGIA
36.0	N	BRAK	NORMA	PRZECIECHŁODZENIE
39.5	T	DUŻA	PRZEKR.	CZARNA OSPA
...	...	...	...	...



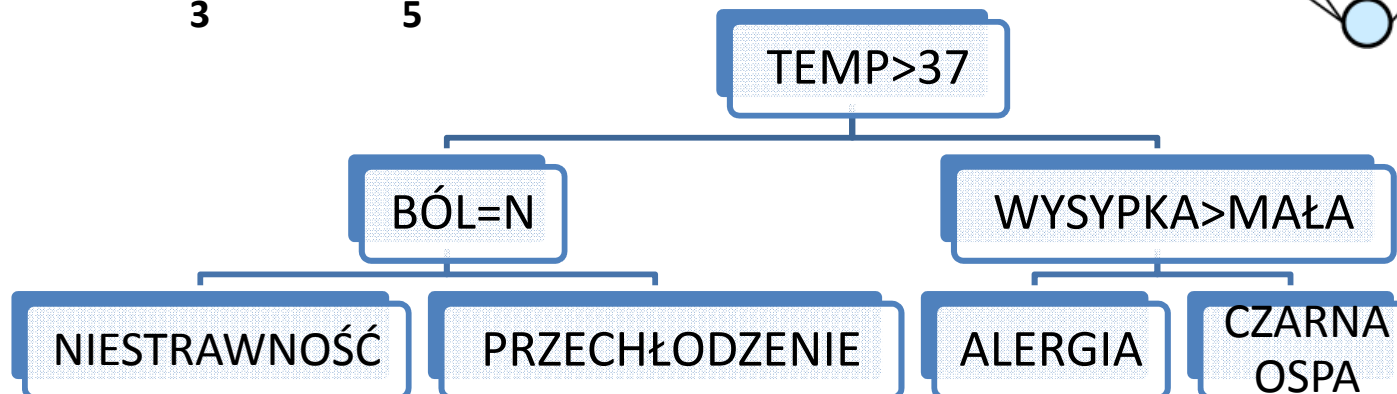
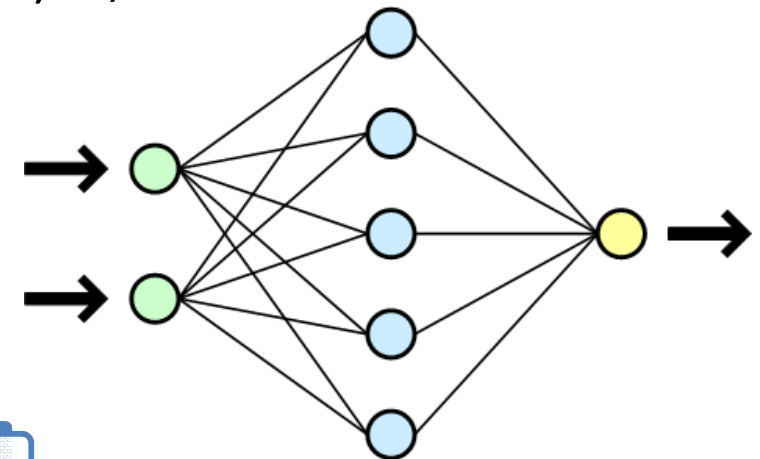
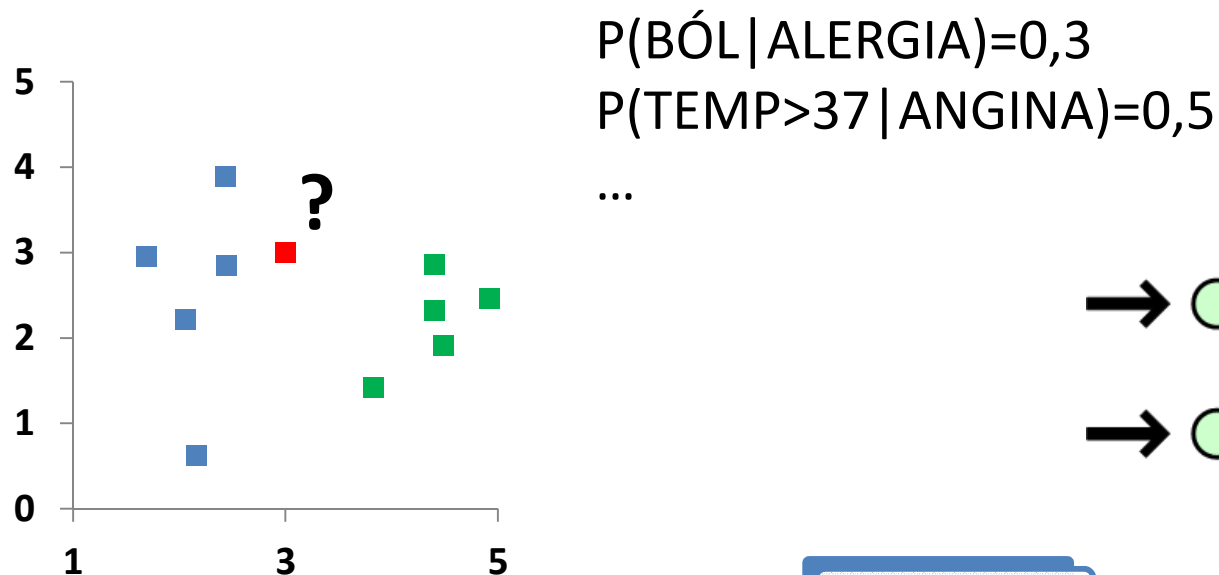
TEMP	BÓL	WYSYPKA	GARDŁO
38.0	N	BRAK	NORMA

DIAGNOZA
ALERGIA

Klasyfikacja (2/2)

gardło = przekrwione $\Rightarrow$ angina

$\neg temp > 37 \wedge ból = N \Rightarrow$ przechłodzenie



Wariancja i odchylenie standardowe

- Miara zmienności zmiennej losowej
 - $\text{var}[X] = E[(X - E[X])^2]$
 - Można estymować wariancję z próby losowej za pomocą wzoru:
 - $s_X^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$, gdzie $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$
- Odchylenie standardowe wyliczane jest jako pierwiastek wariancji.

Kowariancja

- Miara określająca zależność liniową zmiennej losowej
 - $\text{cov}[X, Y] = E[(X - E[X])(Y - E[Y])]$
 - Można estymować wariancję z próby losowej za pomocą wzoru:
 - $q_{X,Y} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$, gdzie $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ (y_i analogicznie).

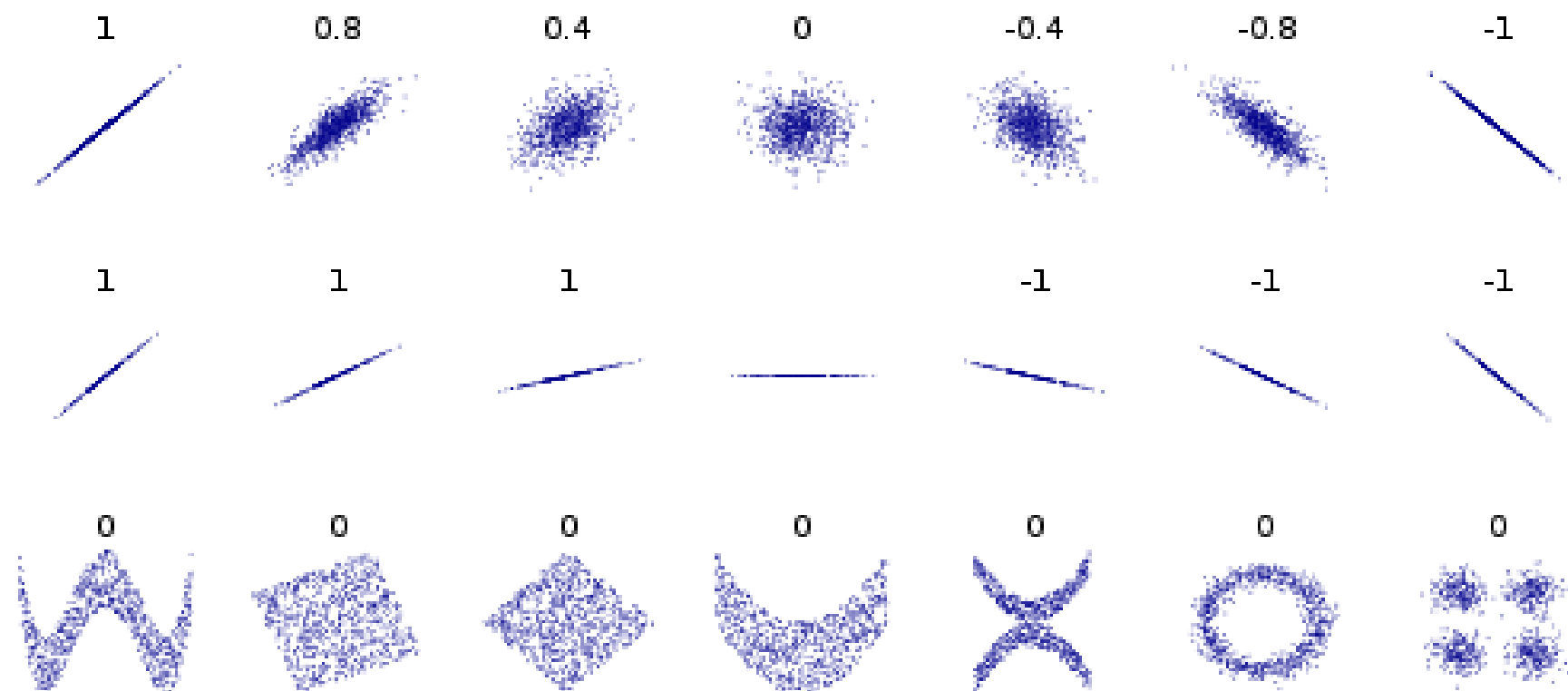
Współczynnik korelacji Pearsona (1)

- Kowariancja może przyjąć dowolne wartości.
- Trudność w analizie.
- Współczynnik korelacji Pearsona, to kowariancja znormalizowana do przedziału $[-1,1]$

- $$\rho = \frac{\text{cov}[X,Y]}{\sqrt{\text{var}[X]}\sqrt{\text{var}[Y]}}$$

- $$r = \frac{q_{x,y}}{s_X s_Y}$$

Współczynnik korelacji Pearsona (2)

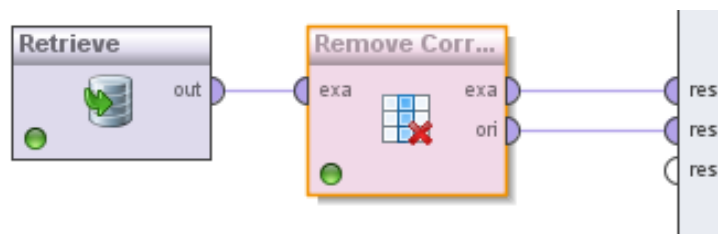


Dataset Sonar

- 60 atrybutów + 1 atrybut decyzyjny
- 208 pomiarów
- Każdy z 60 atrybutów to ilość energii odpowiadającej określonej częstotliwości dźwięku zebranej podczas pomiaru w jakimś okresie czasu.
- Atrybut decyzyjny ma dwie wartości: rock, mine

Usuwanie atrybutów skorelowanych

- Data transformation \ attribute set reduction and transformation \ selection



Wartość graniczna współczynnika Pearsona

Czy ignorujemy znak przed wartością współczynnika Pearsona

The screenshot shows the 'Remove Correlated Attributes' configuration window. It includes the following settings:

- correlation**: 0.7
- filter relation**: greater
- attribute order**: original
- ☒ *use absolute correlation*
- ☐ *use local random seed*

Arrows from the text boxes point to these specific settings: the first box points to the 'correlation' value, the second box points to the 'use absolute correlation' checkbox, and the third box points to the 'filter relation' dropdown.

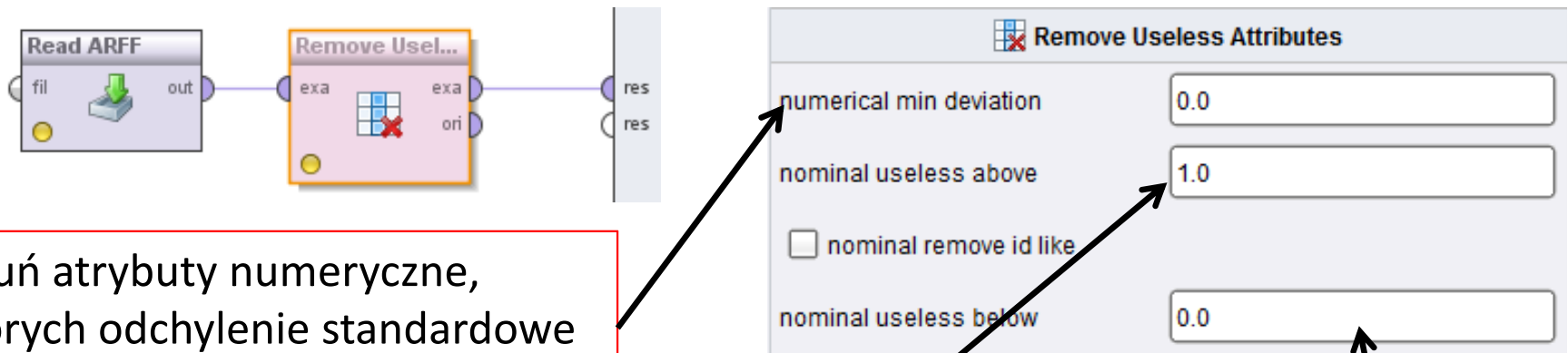
Usuwamy atrybuty o korelacji większej czy mniejszej of wartości granicznej

Zadanie

- Zbuduj przepływ:
 - Źródło danych: sonar
 - Usuń atrybuty o współczynniku korelacji >0.6

Usuwanie atrybutów o małej zmienności

- Data transformation \ attribute set reduction and transformation \ selection



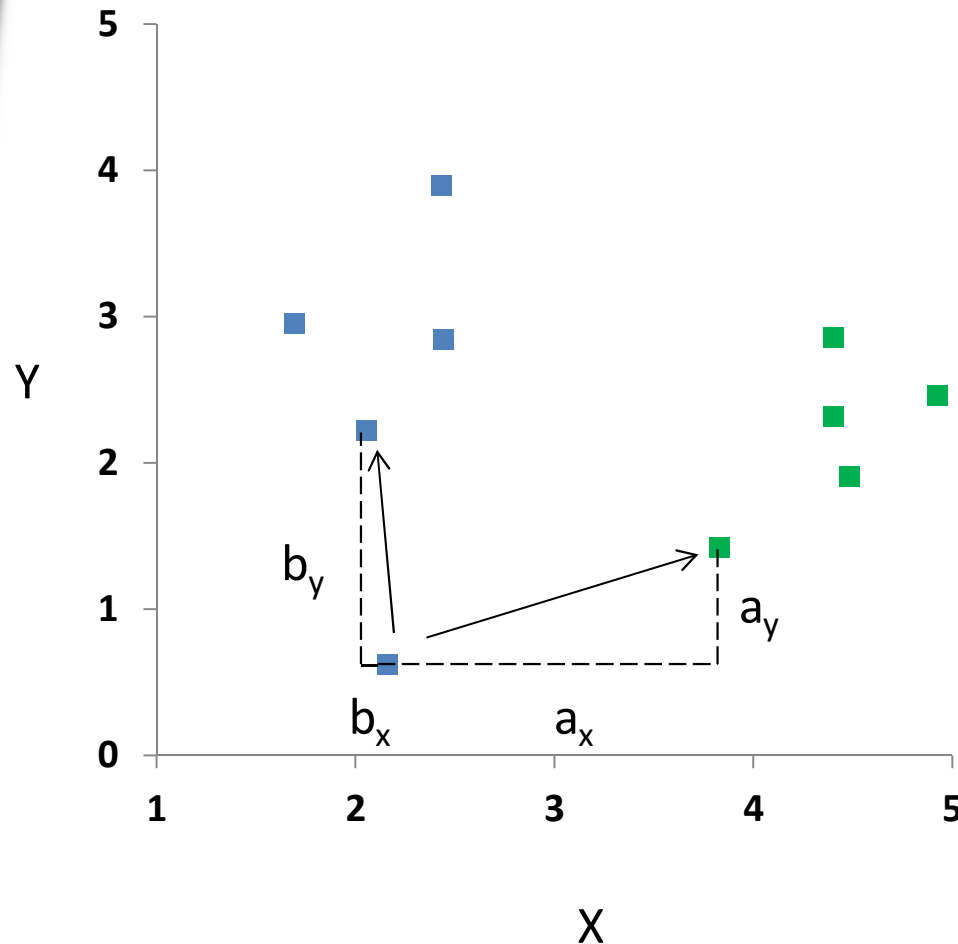
Zadanie

- Zbuduj przepływ:
 - Źródło danych: sonar
 - Usuń atrybuty o odchyleniu standardowym mniejszym niż 0.2

Miara Relief (1)

- Miara Relief to wynik działania algorytmu wyznaczającego relatywną ważność atrybutów.
- Miara relief ocenia jak dobrze poszczególne atrybuty nadają się do przewidywania wartości jednego wybranego atrybutu binarnego, tzw. atrybutu decyzyjnego.

Miara Relief (2)



Na początku:

$$w_x=0$$

$$w_y=0$$

Dla każdego punktu

$$w_x=w_x-b_x^2+a_x^2$$

$$w_y=w_y-b_y^2+a_y^2$$

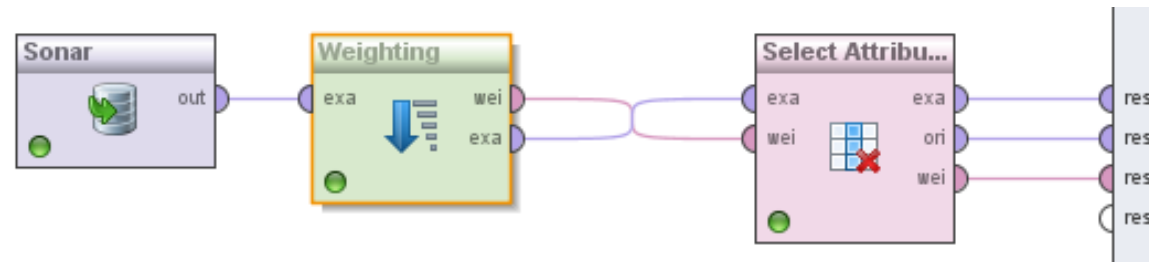
Na końcu

$$w_x=w_x/n$$

$$w_y=w_y/n$$

(n to liczba punktów)

Weight by Relief



Weighting (Weight by Relief)

☐ normalize weights

☒ sort weights

sort direction: ascending

number of neighbors: 10

sample ratio: 1.0

☐ use local random seed

Select Attributes (Select by Weights)

weight relation: greater equals

weight: 0.5

☒ deselect unknown

☒ use absolute weights

Posortuj obliczone wagi

Ile sąsiadów brać pod uwagę dla
każdego obiektu

Jaki procent
obiektów brać pod
uwagę

Pozostaw atrybuty o wagach
które są...

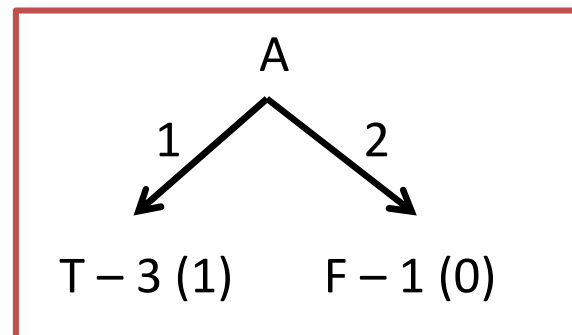
Próg

One rule model (OneR)

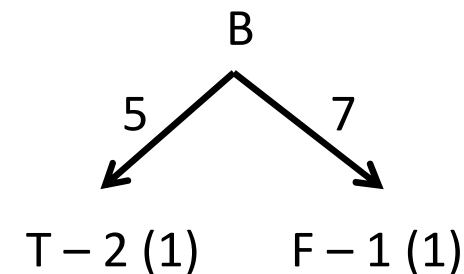
- Bardzo prosty mechanizm klasyfikacji
- Tworzy 1-dno poziomowe drzewo decyzyjne.

A	B	X
1	5	T
1	7	T
1	5	T
1	5	F
2	7	F

		X	
		T	F
A	1	3	1
	2	0	1



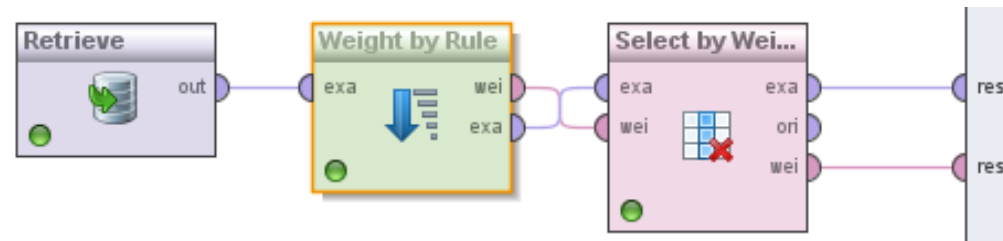
		X	
		T	F
B	5	2	1
	7	1	1



Ocena atrybutu za pomocą OneR

- Usuń oceniany atrybut z tabeli danych.
- Zbuduj model OneR i policz liczbę błędów
- Liczba błędów to waga atrybutu.

Weight by Rule



Weight by Rule

☒ normalize weights

☒ sort weights

sort direction: ascending

Normalizuj wagi do przedziału 0-1

Posortuj obliczone wagi

Select by Weights

weight relation: greater equals

weight: 0.1

☒ deselect unknown

☒ use absolute weights

Pozostaw atrybuty o wagach które są...

Próg

Zadanie 1

- Zbuduj przepływ:
 - Źródło danych: sonar
 - Ocena atrybutów: miara relief
 - Liczba sąsiadów 10,
 - Sample ratio: 1.0 (wykorzystaj wszystkie przykłady)
 - Zachowaj atrybuty o wadze większej niż 0.5

Zadanie 2

- Zmodyfikuj liczbę sąsiadów na:
 - 1
 - 5
 - 15i zobacz jak wpływa to na wyniki (wagi i liczbę pozostawionych atrybutów)
- Zmodyfikuj próg akceptacji na:
 - 0.3
 - 0.5
 - 0.7i zobacz jak to wpływa na wyniki (liczbę pozostawionych atrybutów)

Zadanie 3

- Zbuduj przepływ:
 - Źródło danych: sonar
 - Ocena atrybutów: miara rule
 - Obejrzyj wynikowe wagi
 - Zachowaj atrybuty o wadze większej niż:
 - 0.5
 - 0.7
 - 0.9
 - 0.98
 - Zobacz jak wpływa to na liczbę pozostawionych atrybutów