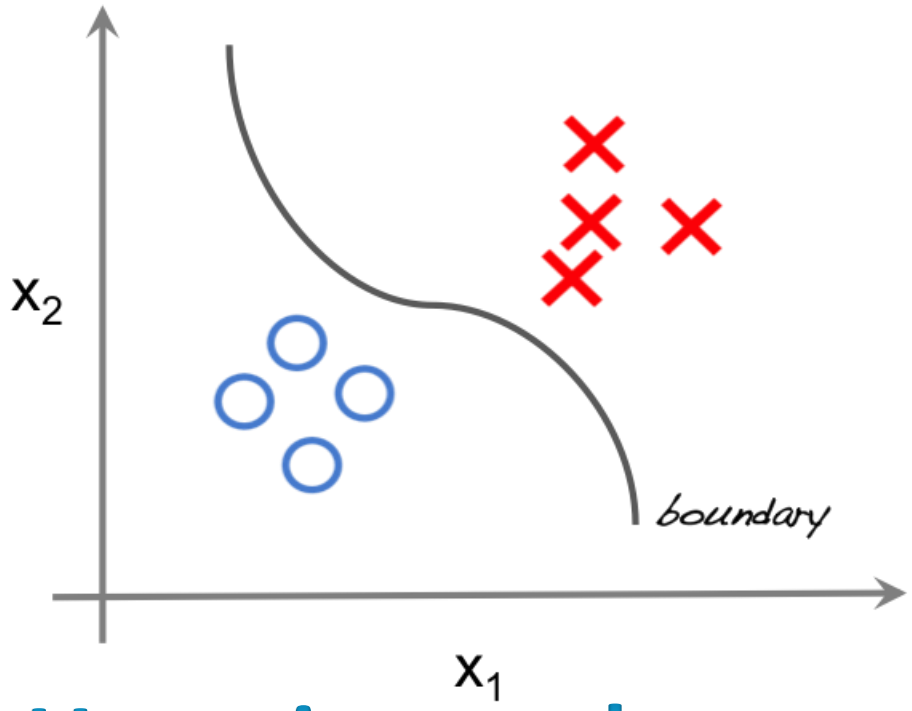
Uczenie maszynowe

ZADANIE KLASYFIKACJI CZ. 1

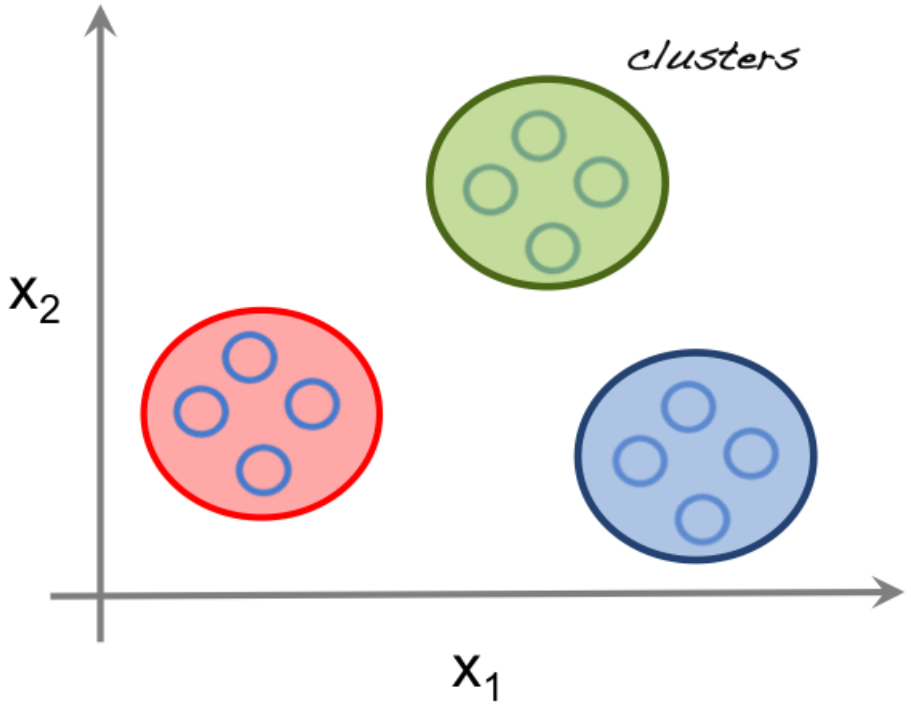
Typy uczenia

Supervised learning



Uczenie z nadzorem

Unsupervised learning



Uczenie bez nadzoru

Uczenie z nadzorem – zadanie klasyfikacji

Przykład: klasyfikacja kosaćców

Iris setosa
(kosaciec szczecinkowy)

Iris virginica

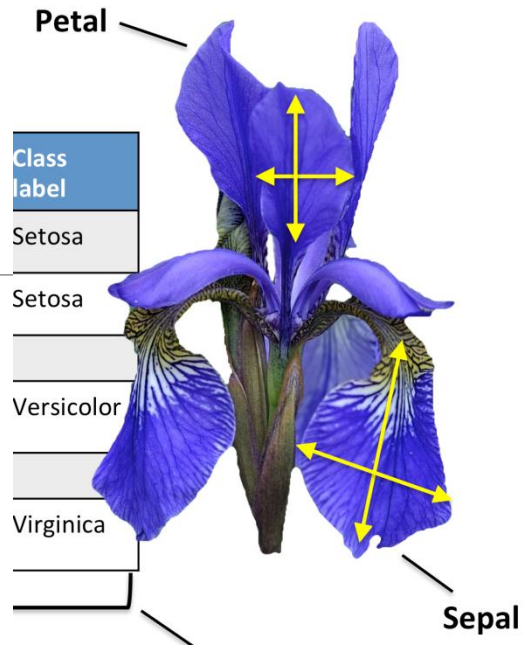
Iris versicolor
(kosaciec różnobarwny)

Zbiór - IRIS

Atrybuty **Etykiety**

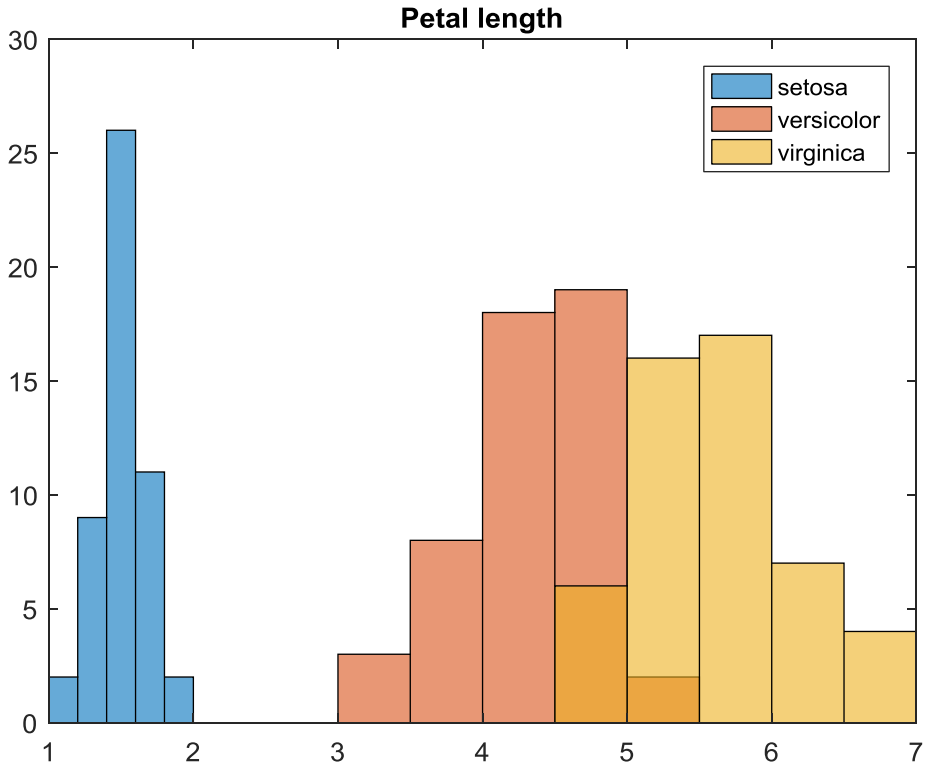
Instancje

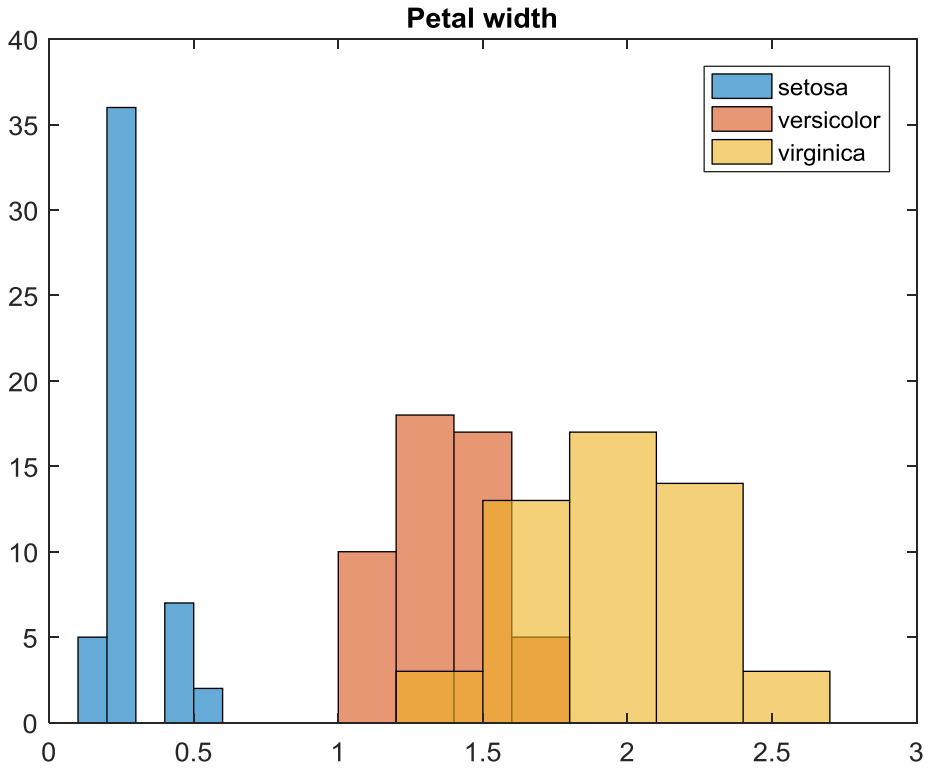
SL	SW	PL	PW	Klasa
5.1	3.5	1.4	0.2	Iris-setosa
4.9	3.0	1.4	0.2	Iris-setosa
4.7	3.2	1.3	0.2	Iris-setosa
7.7	3.8	6.7	2.2	Iris-virginica
7.7	2.6	6.9	2.3	Iris-virginica
6.0	2.2	5.0	1.5	Iris-virginica
6.1	2.9	4.7	1.4	Iris-versicolor
5.6	2.9	3.6	1.3	Iris-versicolor
6.7	3.1	4.4	1.4	Iris-versicolor



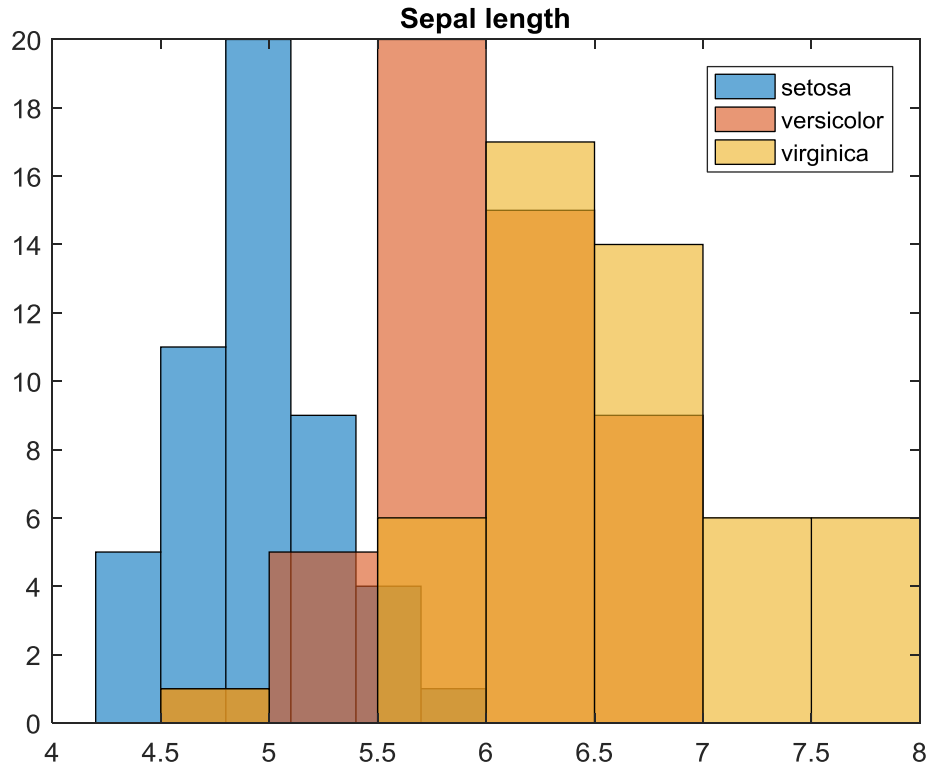
SL (sepal length) - długość działki kielicha kwiatu (w cm);
SW (sepal width) - szerokość działki kielicha (cm);
PL (petal length) - długość płatka (w cm);
PW (petal width) - szerokość płatka (w cm)

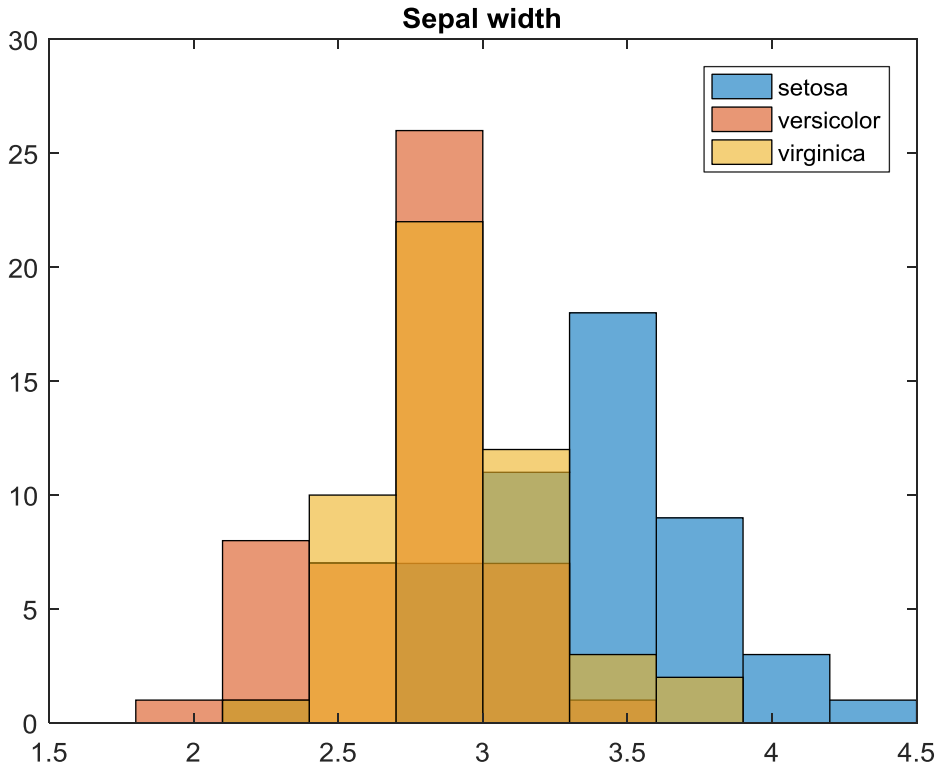
Histogramy dla zbioru Iris 1/2



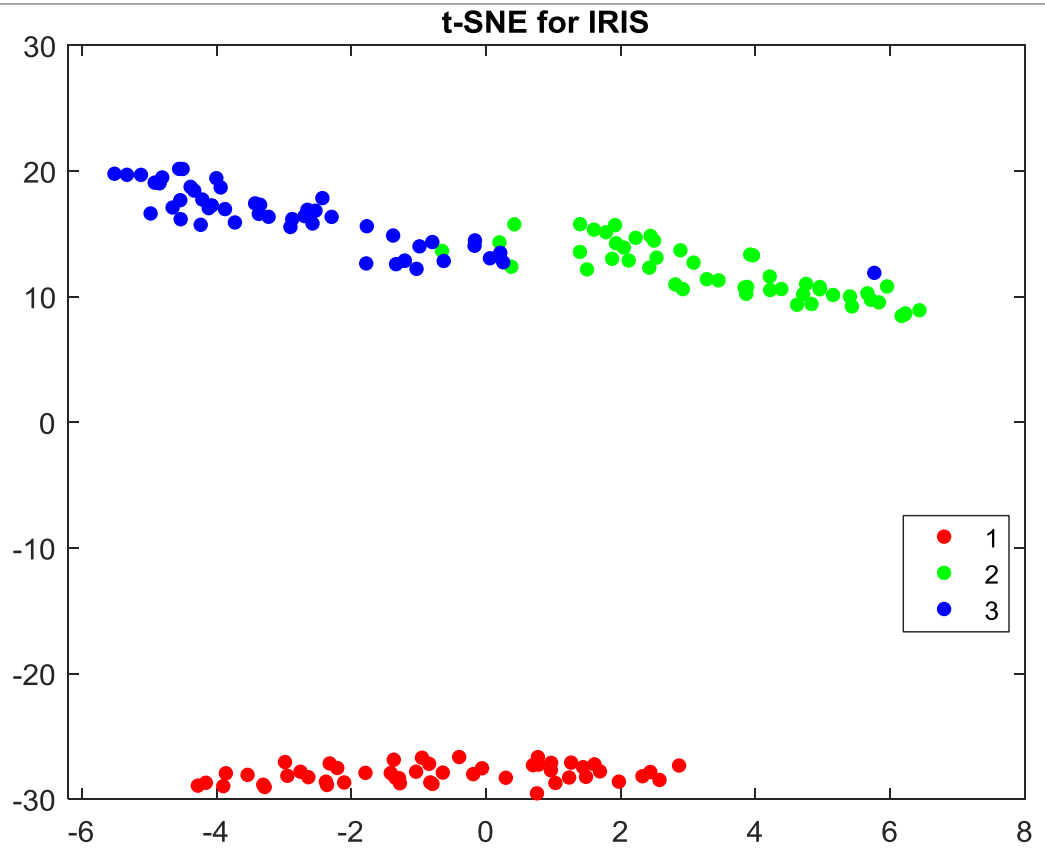


Histogramy dla zbioru Iris 2/2



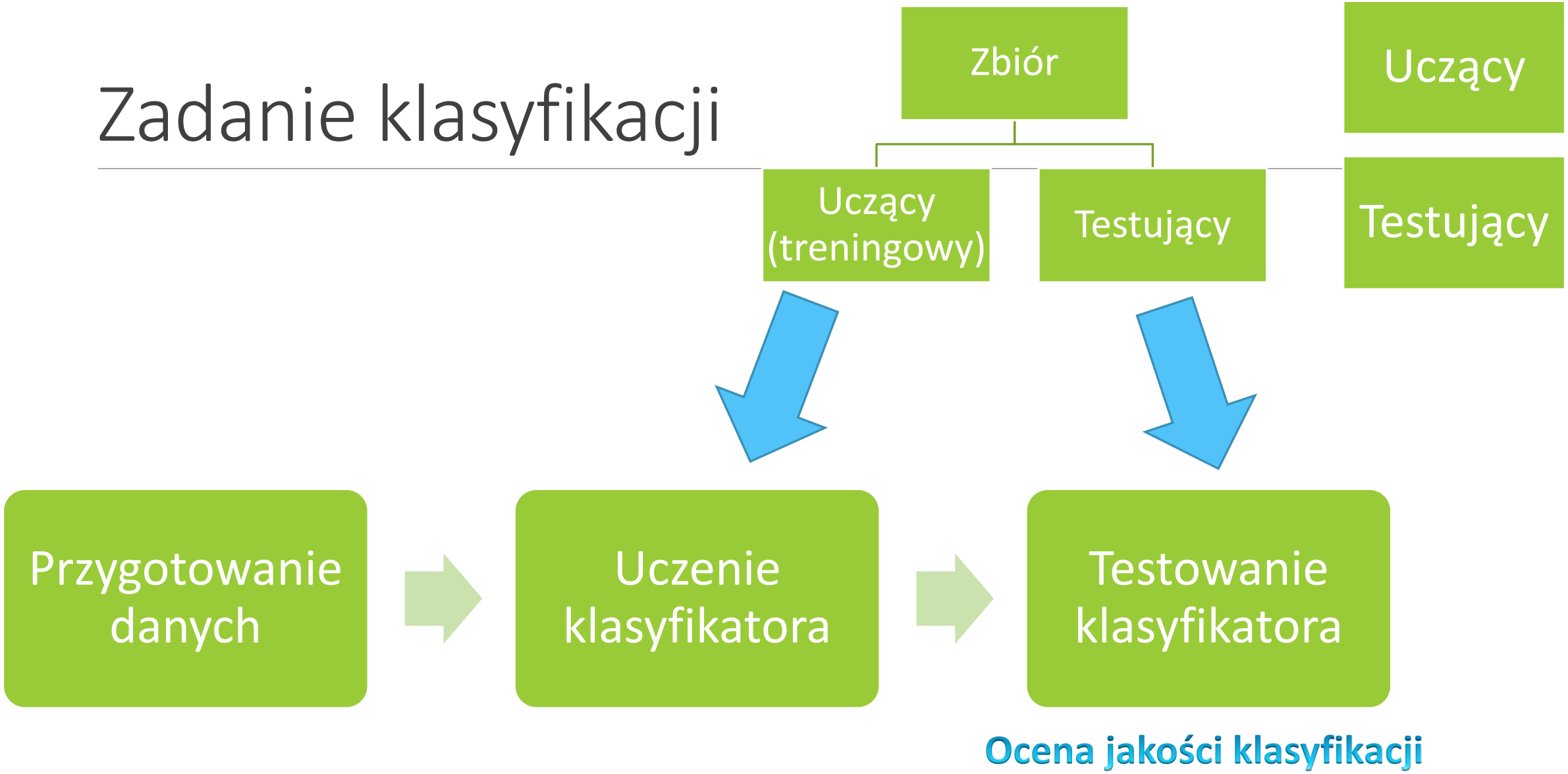


Wizualizacja tSNE dla zbioru Iris



Legenda: 1 – setosa, 2 – versicolor, 3 - virginica

Zadanie klasyfikacji



Klasyfikator kNN (*k–nearest neighbour classifier*)

kNN– klasyfikator *k*-najbliższych sąsiadów

Pochodzenie nazwy metody: poszukiwanie przez algorytm *k* przypadków w najbliższym sąsiedztwie nowego punktu.

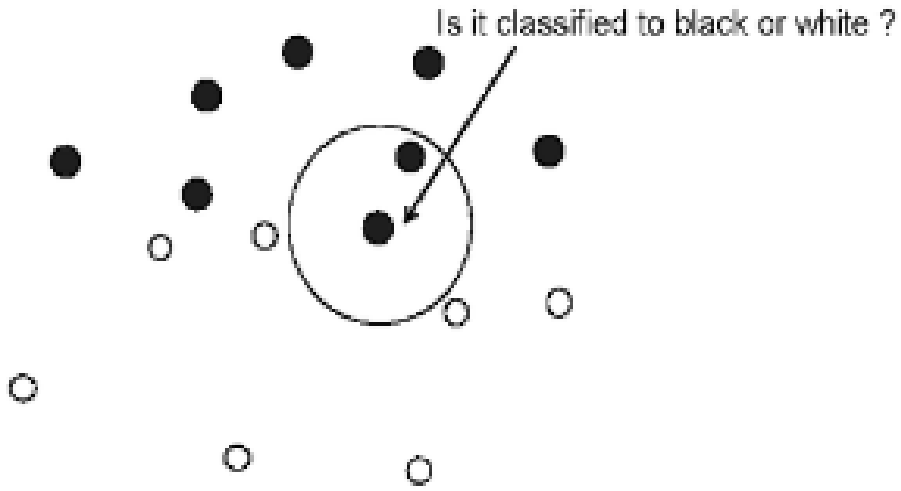
Zadanie klasyfikacji: przypisanie etykiety do rozpatrywanego przypadku, znając jego *k*-najbliższych sąsiadów w przestrzeni.

Dobór liczby *k:* na tyle duże, by minimalizować prawdopodobieństwo błędnych klasyfikacji i na tyle małe, aby odnaleźć dostatecznie bliskich sąsiadów nowego punktu.

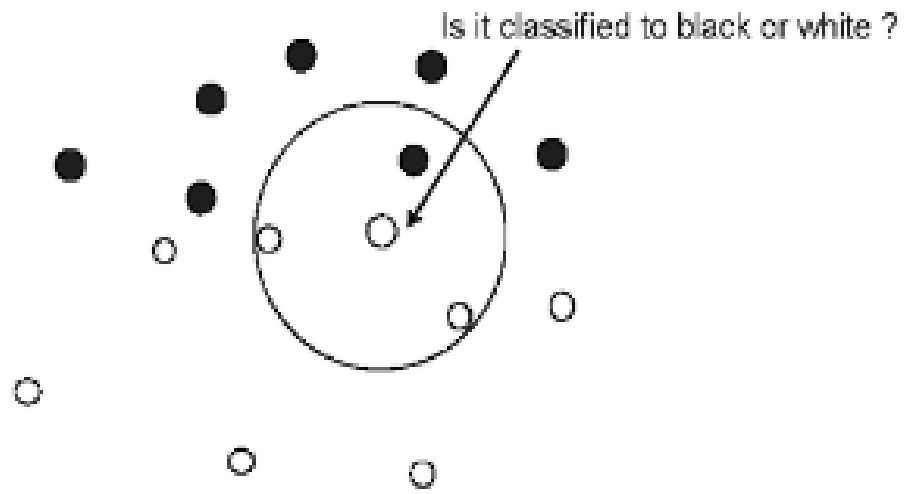
kNN – leniwy klasyfikator….

Zasada działania kNN

1-Nearest Neighbor



3-Nearest Neighbor



http://beta.cambridgespark.com/courses/jpm/01-module.html

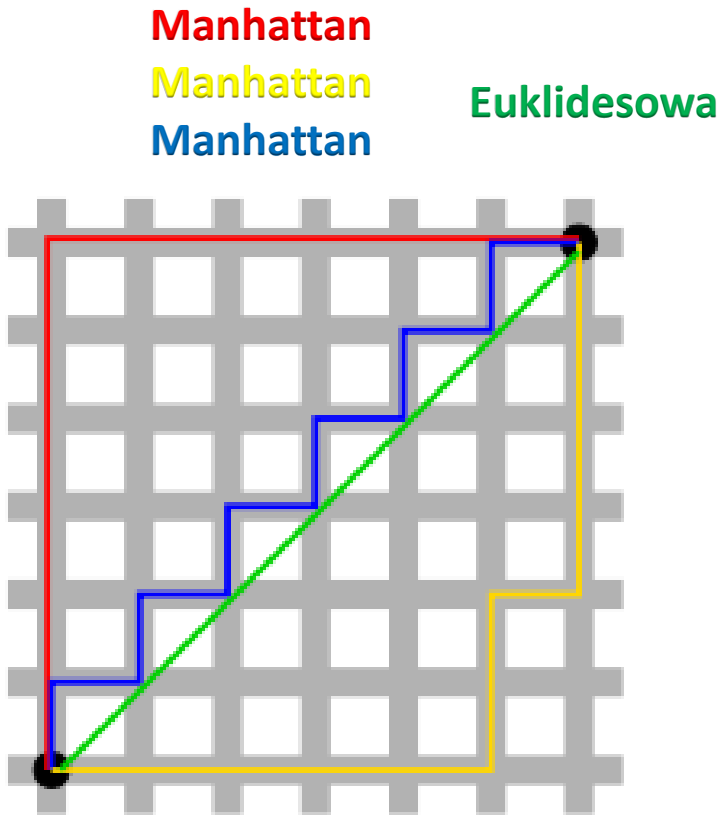
Miary odległości między punktami **x** i **y**

Miara euklidesowa

$$d_e(\mathbf{x},\mathbf{y}) = \sqrt{(y_1 - x_1)^2 + \cdots + (y_n - x_n)^2}$$

Miara Manhattan (znana również jako miejska, taksówkowa)

$$d_m(\mathbf{x},\mathbf{y}) = \sum_{k=1}^{n} |x_k - y_k|$$



Porównanie metryki Manhattan z euklidesową

https://pl.wikipedia.org/wiki/Przestrze%C5%84_metryczna

Głosowanie - klasyfikacja nowego przypadku

- Większościowe równoprawne - punkt przypisywany jest do klasy o największej liczności
- Ważone odległością - wartości wag obliczane są jako odwrotność odległości między szukanym punktem a najbliższymi sąsiadami:

$$w(x, p_i) = \frac{1}{d_{xp_i}},$$

gdzie: d_{xpi} - odległość między punktem x a i-tym punktem przykładowym. Dla każdej z klas sumuje się wagi i klasyfikowanemu przykładowi przypisuje się klasę z najwyższą sumą.

- Ważone kwadratem odległości - wartości wag obliczane są jako odwrotność kwadratu odległości między szukanym punktem a najbliższymi sąsiadami:

$$w(x, p_i) = \frac{1}{{d_{xp_i}}^2},$$

gdzie: d_{xpi} - odległość między punktem x a i-tym punktem przykładowym. Dla każdej z klas sumuje się wagi i klasyfikowanemu przykładowi przypisuje się klasę z najwyższą sumą.

Problem: parzysta liczba sąsiadów ….

- Losowe przyporządkowanie do klasy
- Zmniejszanie/zwiększanie liczby k aż do ustalenia „zwycięzcy”
- Ważenie odległości
- …

Problem: czułość na strukturę danych

Rozwiązanie - standaryzacja atrybutów (zmienna uzyskuje średnią wartość oczekiwaną zero i odchylenie standardowe jeden)

Standaryzacja Z (najczęstsza):

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

gdzie: x – zmienna niestandaryzowana, μ – średnia z populacji, σ – odchylenie standardowe populacji.

https://pl.wikipedia.org/wiki/Standaryzacja_(statystyka)

Miary jakości klasyfikacji

Macierz pomyłek (*confusion matrix*)

W zależności od działania klasyfikatora wyróżnia się cztery przypadki:
◦ liczba prawdziwie rozpoznanych przypadków pozytywnych (ang. *True Positive* - TP),
◦ liczba nieprawdziwie rozpoznanych przypadków pozytywnych (ang. *False Positive* - FP),
◦ liczba nieprawdziwie rozpoznanych przypadków negatywnych (ang. *False Negative* - FN),
◦ liczba prawdziwie rozpoznanych przypadków negatywnych (ang. *True Negative* - TN).

		Prawdziwa klasyfikacja	
		Klasa pozytywna	Klasa negatywna
Wynik klasyfikacji	Klasa pozytywna	TP	FP
	Klasa negatywna	FN	TN

Macierz pomyłek - przykład

Klasyfikacja wiadomości e-mail - SPAM i dobre wiadomości

Klasa pozytywna: SPAM

Dane wejściowe: 37 SPAM, 63 dobrych

Wynik klasyfikacji: 33 wiadomości uznanych za SPAM (w tym 27 to rzeczywisty SPAM), 67 wiadomości uznanych za dobre (w tym 57 rzeczywiście dobrych)

SPAM Dobre

Wynik klasyfikacji	Prawdziwa klasyfikacja: Klasa pozytywna	Prawdziwa klasyfikacja: Klasa negatywna
Klasa pozytywna (SPAM)	TP = 27	FP = 6
Klasa negatywna (Dobre)	FN = 10	TN = 57

Pomysł na przykład: http://aimotion.blogspot.com/2010/08/tools-for-machine-learning-performance.html

Miary jakości klasyfikacji

Dokładność klasyfikacji (ang. accuracy - ACC) określa procent poprawnie sklasyfikowanych przypadków:

$$ACC\% = \frac{TP + TN}{TP + TN + FP + FN} \cdot 100\%$$

Specyficzność (ang. True Negative Rate - TNR lub specificity) określa stosunek poprawnie rozpoznanych przypadków negatywnych do liczby wszystkich przypadków negatywnych:

$$TNR = 1 - FPR = 1 - \frac{FP}{FP + TN} = \frac{TN}{TN + FP}$$

Czułość (ang. True Positive Rate - TPR lub sensitivity), definiowana również jako miara recall określa stosunek poprawnie rozpoznanych przypadków pozytywnych do liczby wszystkich przypadków pozytywnych:

$$TPR = \frac{TP}{TP + FN}$$

Miary jakości klasyfikacji SPAM

$$ACC\% = \frac{TP + TN}{TP + TN + FP + FN} \cdot 100\%$$

ACC = 84/100 = 0,84

$$TNR = \frac{TN}{TN + FP}$$

TNR = 57/(57+6) ≈ 0,90

$$TPR = \frac{TP}{TP + FN}$$

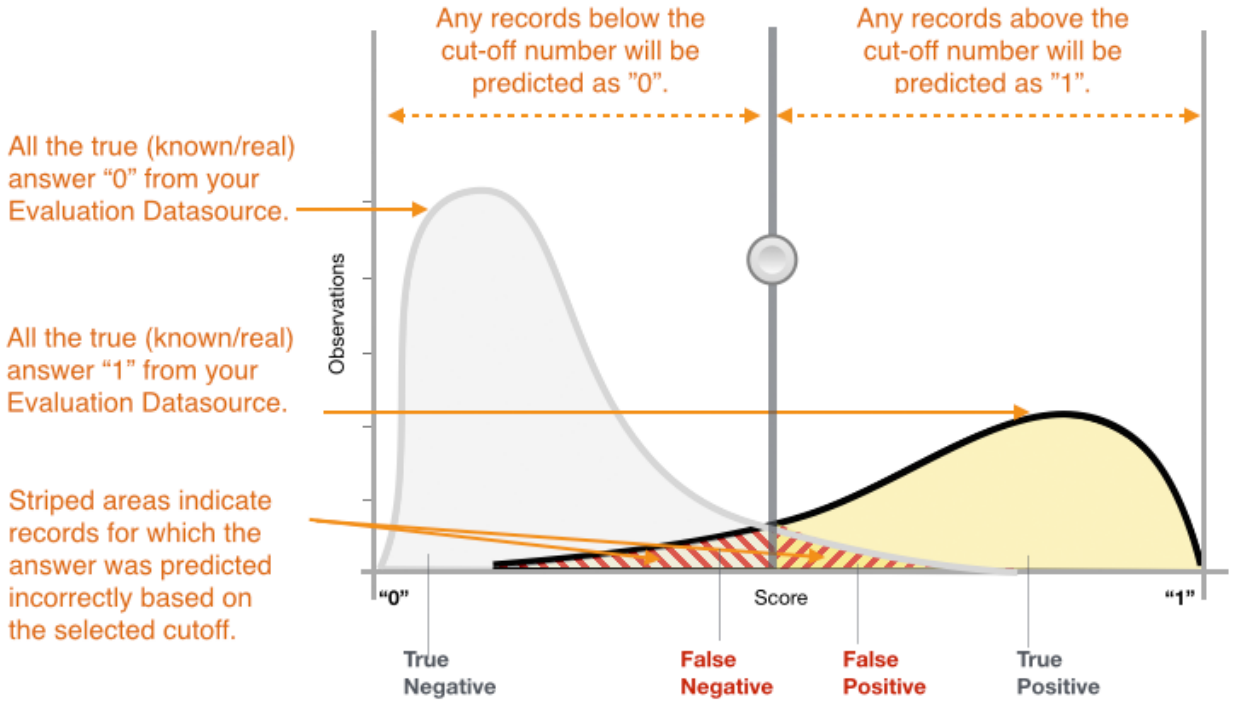
TPR = 27/(27+10) ≈ 0,73

		SPAM Prawdziwa klasyfikacja Dobre	
	SPAM	Klasa pozytywna	Klasa negatywna
Wynik klasyfikacji	Klasa pozytywna	TP = 27	FP = 6
	Klasa negatywna	FN = 10	TN = 57
	Dobre		

Nakładanie się rozkładów dwóch klas

Nakładanie się rozkładów dwóch klas

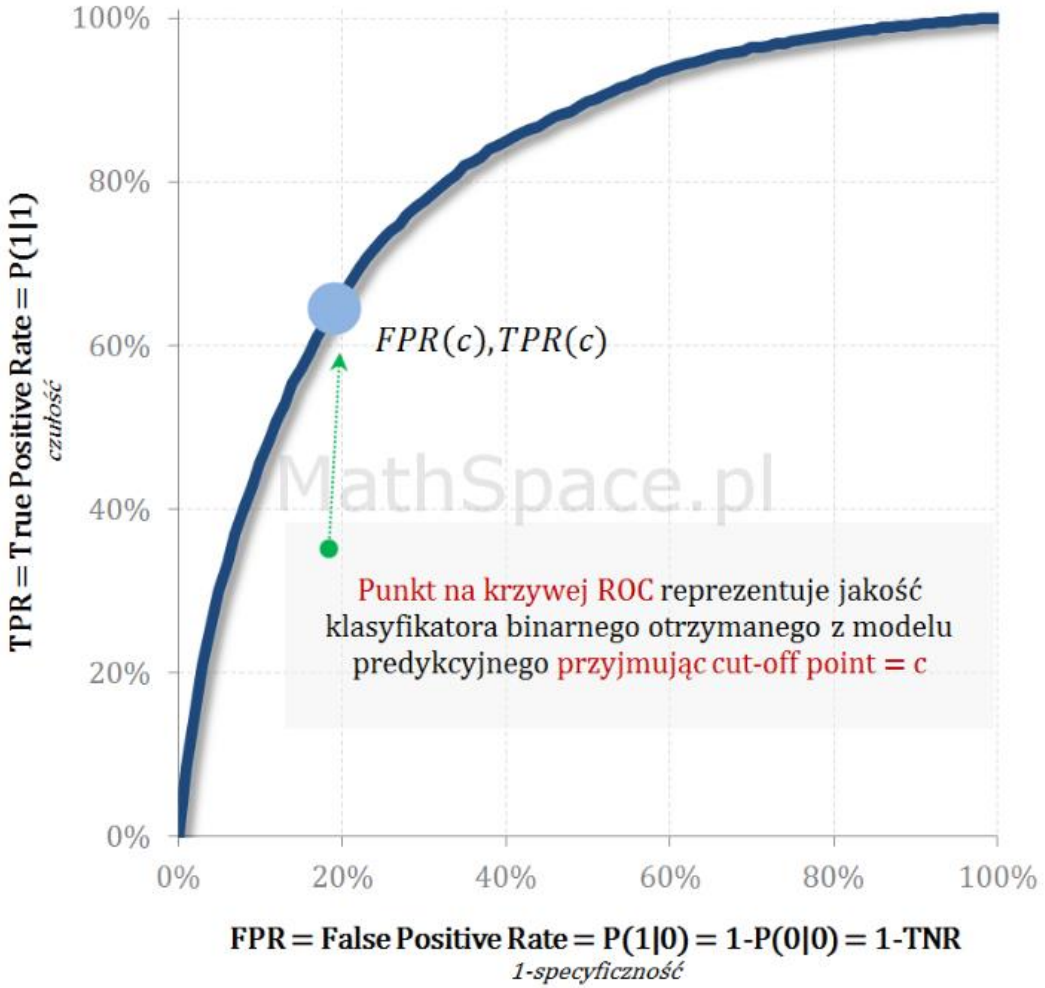
Rozwiązanie: wybór progu/punktu odcięcia (*cutpoint*)



http://mathspace.pl/matematyka/receiver-operating-characteristic-krzywa-roc-czyli-ocena-jakosci-klasyfikacji-czesc-7/

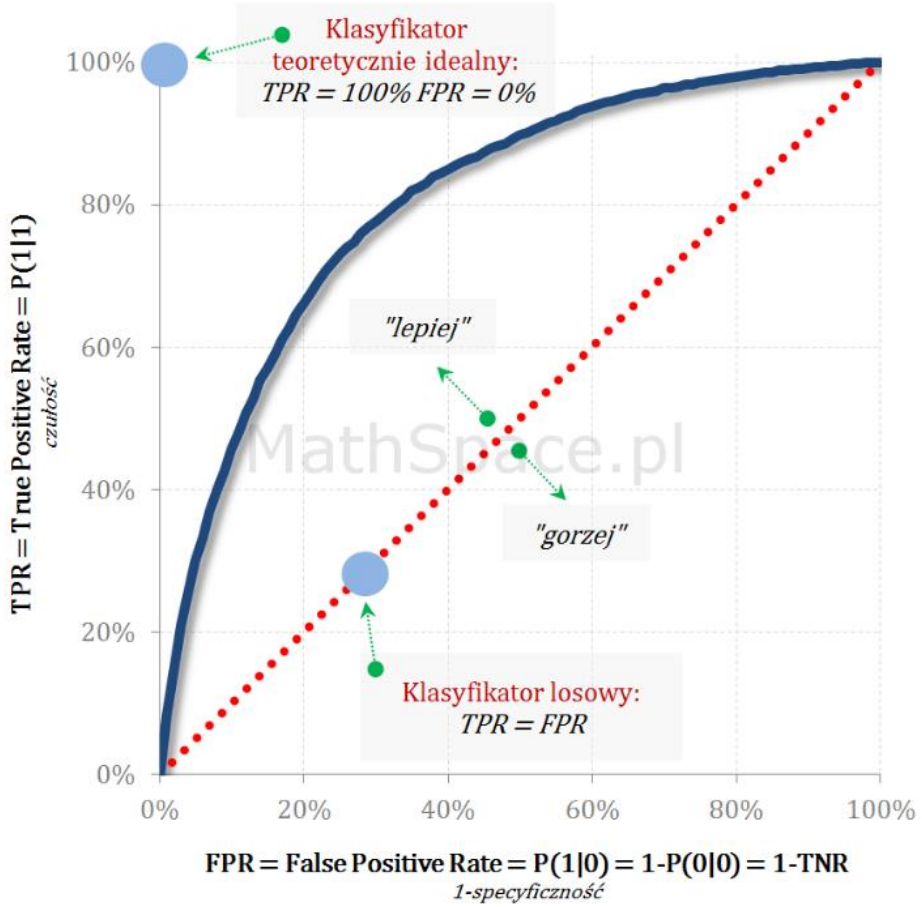
Krzywa ROC (*Receiver Operating Characteristic*)

- Odróżnienie sygnału będącego informacją (np. sygnały z maszyn, organizmów żywych) od wzorców przypadkowych nie zawierających informacji (szum, tło, aktywność losowa)
- Statystyka: „*Krzywa ROC jest graficzną reprezentacją efektywności modelu predykcyjnego poprzez wykreślenie charakterystyki jakościowej klasyfikatorów binarnych powstałych z modelu przy zastosowaniu wielu różnych punktów odcięcia.*”



http://mathspace.pl/matematyka/receiver-operating-characteristic-krzywa-roc-czyli-ocena-jakosci-klasyfikacji-czesc-7/

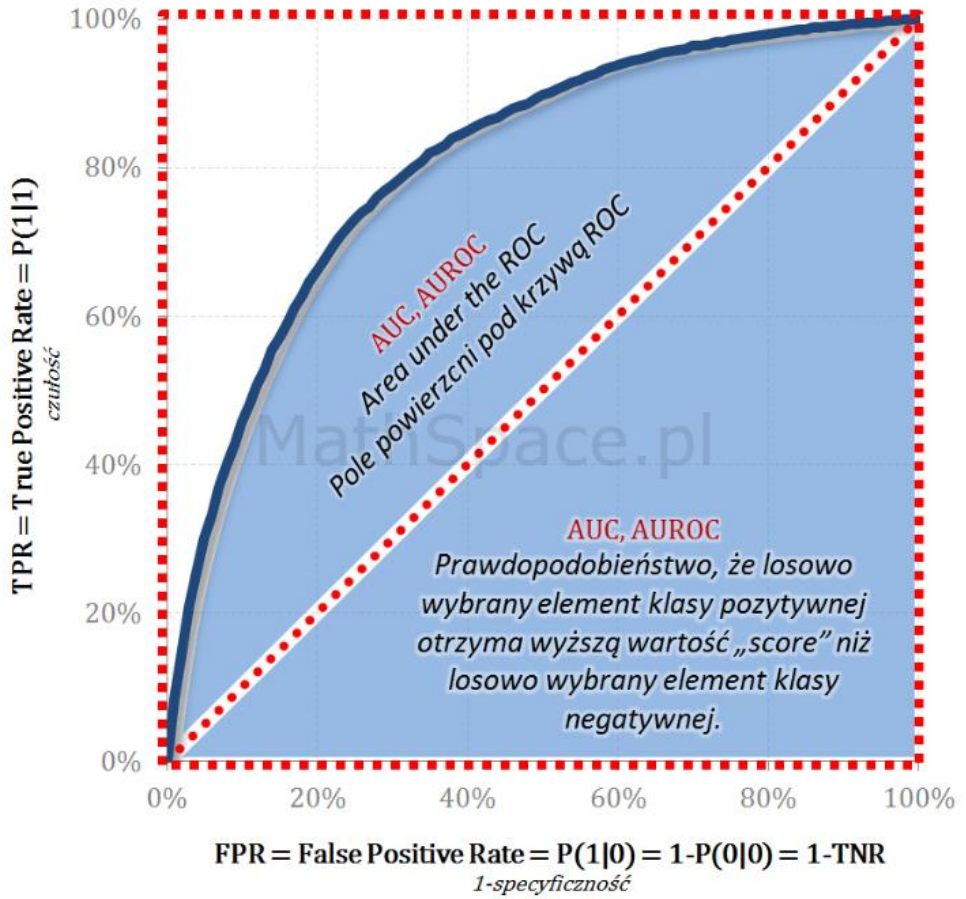
Krzywa ROC



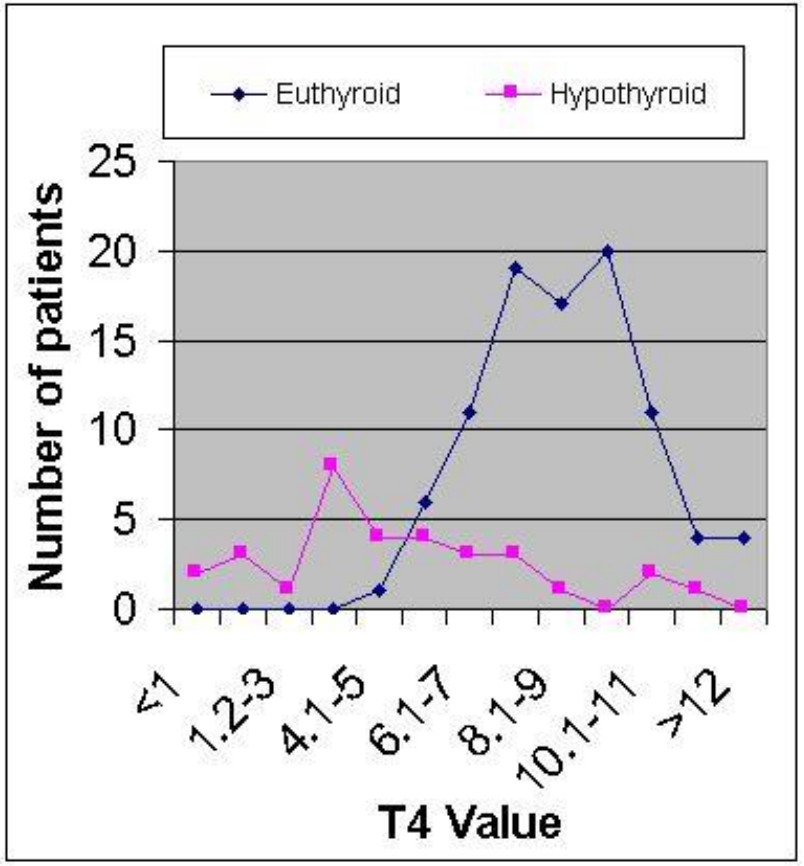
http://mathspace.pl/matematyka/receiver-operating-characteristic-krzywa-roc-czyli-ocena-jakosci-klasyfikacji-czesc-7/

AUROC (Area Under the ROC)

- Całość pola powierzchni pod krzywą ROC w odniesieniu do pola idealnego modelu (pola kwadratu o boku 1)
- Interpretacja: „*prawdopodobieństwo, że badany model predykcyjny oceni wyżej losowy element klasy pozytywnej od losowego elementu klasy negatywnej.”*



ROC przykład – choroby tarczycy



Próg odcięcia T4 = 5

T4 value	Hypothyroid	Euthyroid
5 or less	18	1
> 5	14	92
Totals:	32	93

$$TNR = \frac{TN}{TN + FP}$$

TNR = 92/(92+1) ≈ 0,99

$$TPR = \frac{TP}{TP + FN}$$

TPR = 18/(18+14) ≈ 0,56

http://gim.unmc.edu/dxtests/ROC1.htm

$$TNR = \frac{TN}{TN + FP}$$

$$TPR = \frac{TP}{TP + FN}$$

ROC przykład – choroby tarczycy

Próg odcięcia T4 = 7

T4 value	Hypothyroid	Euthyroid
7 or less	25	18
> 7	7	75
Totals:	32	93

TNR = 75/(75+18) ≈ 0,81

TPR = 25/(25+7) ≈ 0,78

Próg odcięcia T4 = 9

T4 value	Hypothyroid	Euthyroid
< 9	29	54
9 or more	3	39
Totals:	32	93

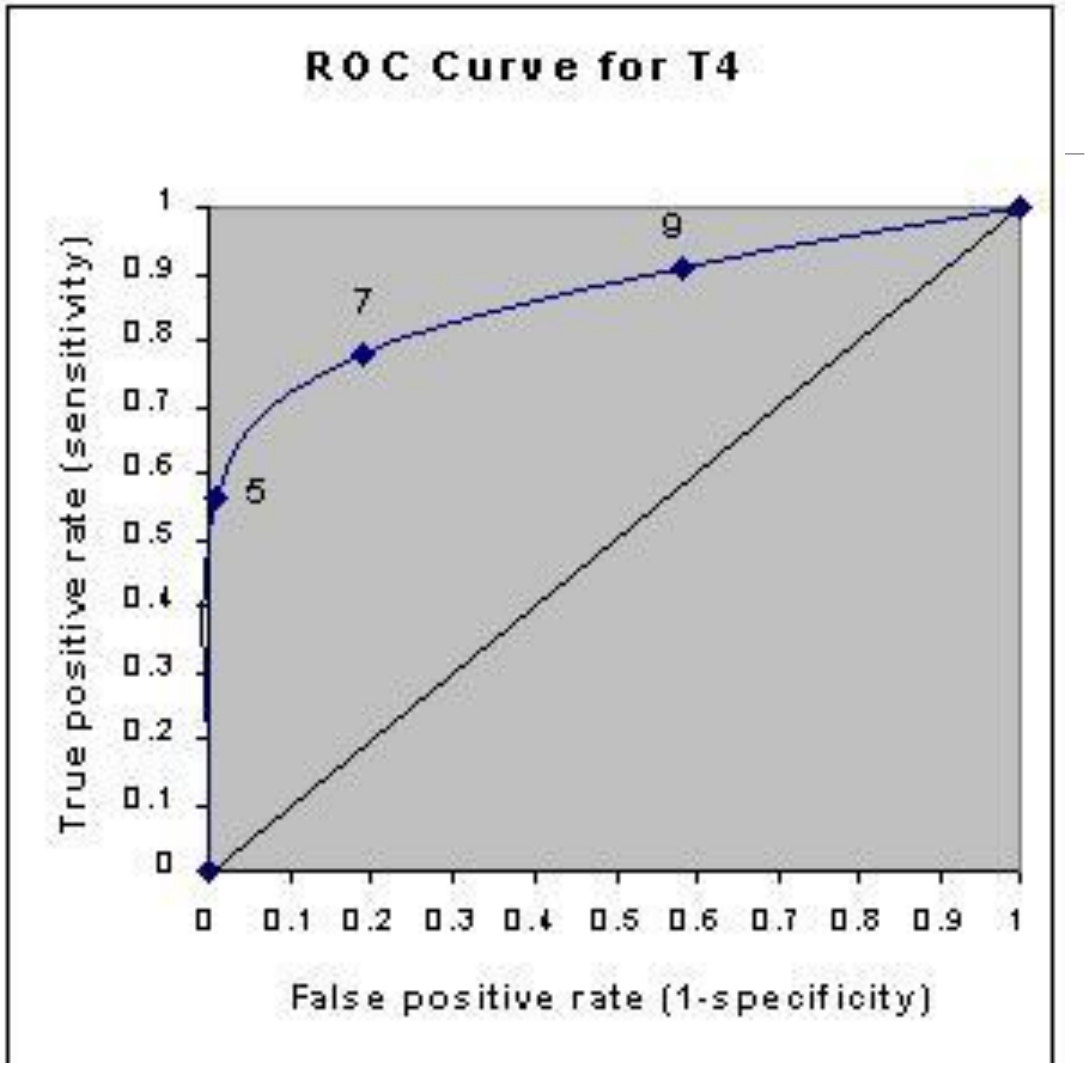
TNR = 39/(39+54) ≈ 0,42

TPR = 29/(29+3) ≈ 0,91

http://gim.unmc.edu/dxtests/ROC1.htm

ROC przykład

Odcięcie	TPR	TNR	1-TNR
5	0,56	0,99	0,01
7	0,78	0,81	0,19
9	0,91	0,42	0,58



http://gim.unmc.edu/dxtests/ROC1.htm