GDAŃSK UNIVERSITY OF TECHNOLOGY | DOCTORAL SCHOOL

RESEARCH UNIVERSITY EXCELLENCE INITIATIVE

Miary i metryki w klasyfikacji i transkrypcji

mgr inż. Tomasz Piernicki/dr inż. A. Kurowski

Wprowadzenie do Sztucznej Inteligencji

mgr inż. Tomasz Piernicki/

dr inż. A. Kurowski

LAF/KSM WETI

Wprowadzenie

Niezależnie od wybranej techniki uczenia maszynowego po wytrenowaniu modelu zwykle pojawia się konieczność przeprowadzenia bardziej rozległego testu skuteczności wybranego rozwiązania.

Może być konieczne przetestowanie, jak dana technika uczenia maszynowego sprawdza się w kontekście danego typu zbioru.

Często konieczna jest też wiedza dotycząca tego, jak w ocenie jakości działania algorytmu uwzględnić fakt, że dysponujemy zbiorem danych o nierównych liczebnościach klas.

Tego typu zagadnienia wymagają szerszego podejścia do analizy skuteczności modelu uczenia maszynowego niż prosty podział danych treningowych na trzy podzbiory.

Wprowadzenie

Przykładowe sytuacje, w których konieczne jest specyficzne podejście do oceny działania algorytmu:

• przygotowywany jest algorytm wykonujący zadanie klasyfikacji i liczba przykładów nie jest podobna dla wszystkich klas,

• przygotowywany jest algorytm wykonujący zadanie regresji i liczba przykładów nie jest równomierna dla całej dziedziny, dla której obliczana jest regresja,

• musimy wybrać lepszy z dwóch modeli, które nieznacznie różnią się osiąganą jakością klasyfikacji/regresji,

• chcemy przewidzieć, jak dany model zachowa się po napotkaniu różnorodnych danych – na przykład po to by przygotować się na zjawisko niedopasowania modelu do danych rzeczywistych (ang. data shift),

• przygotować się na fakt, że model będzie dotrenowywany w trakcie eksploatacji i chcemy, aby osiągane efekty treningu nie miały dużego rozrzutu jakości i były zawsze możliwie wysokie.

Wprowadzenie

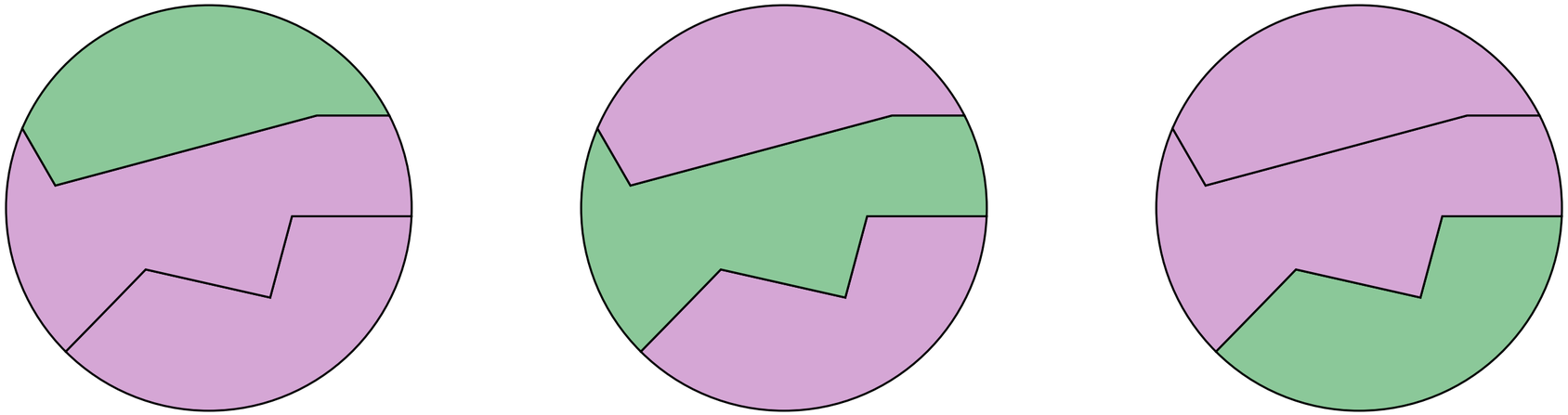
Jeżeli przygotowujemy duży, długo uczący się model, to mamy ograniczoną możliwość testowania jakości modelu poprzez jego wielokrotny trening. W takim przypadku pomocne jest:
• testowanie modelu za pomocą **kilku różnych zbiorów**,
• a także poprzez **testowanie na zbiorze podzielonym na kilka mniejszych** (równolicznych, ale często mogących mieć części wspólne) zbiorów.

Jeżeli mamy model, który możemy trenować wielokrotnie, to możemy posłużyć się następującymi metodami:

• ***k*-krotna walidacja krzyżowa** (*k-fold cross-validation*) – podział zbioru na *k* **rozłącznych, równolicznych** podzbiorów, trening modelu na każdym możliwym zestawie k-1 podzbiorów i test na jednym pozostawionym podzbiorze.

• **procedura 5x2CV** – 5-krotne powtórzenie procedury 2-krotnej walidacji – jest to optymalny ze statystycznego punktu widzenia dobór wartości *k*, oraz liczby powtórzeń procedury walidacji krzyżowej.

Przykład: 3-krotna walidacja krzyżowa

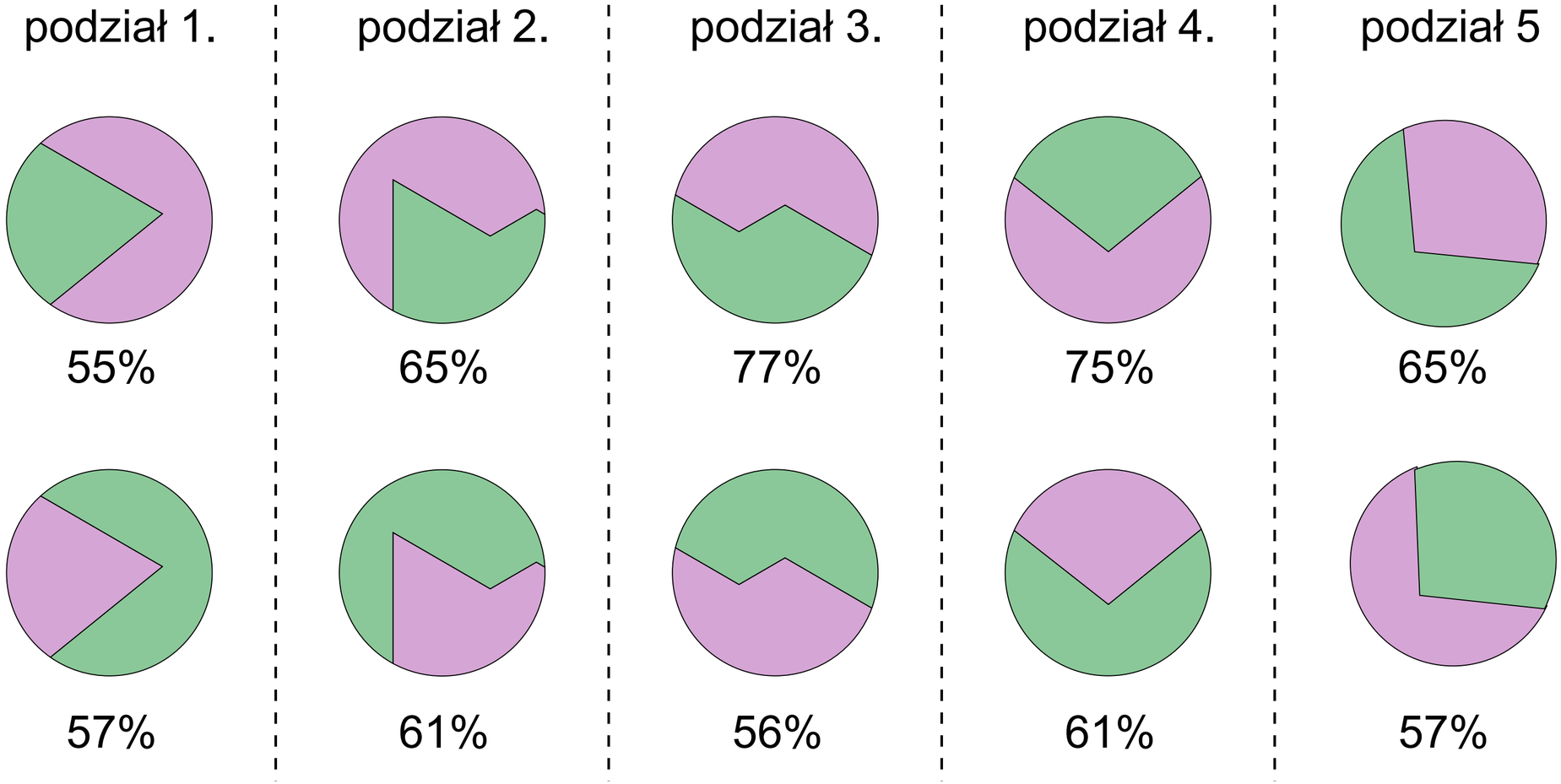


83% 91% 88%

zbiór treningowy
zbiór walidacyjny

wartościami procentowymi oznaczone są skuteczności klasyfikacji osiągnięte dla każdego sposobu przypisania zbiorów testowych i walidacyjnych

Przykład: procedura 5x2CV



zbiór treningowy
zbiór walidacyjny

wartościami procentowymi oznaczone są skuteczności klasyfikacji osiągnięte dla każdego sposobu przypisania zbiorów testowych i walidacyjnych

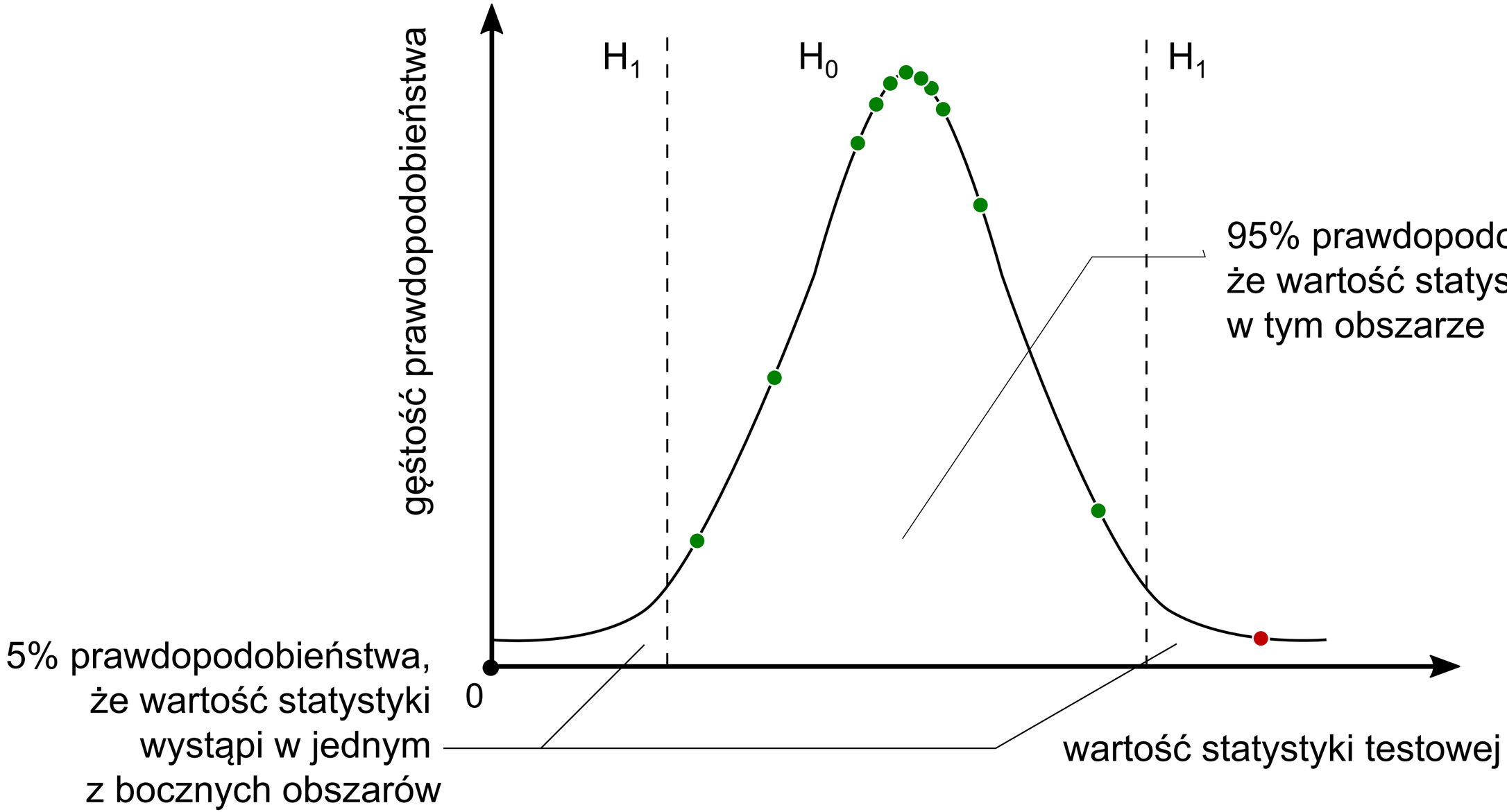
Analiza danych z procedur bazujących na walidacji krzyżowej

Walidacja krzyżowa i procedura 5x2CV dostarczają informacji nie tylko o dokładności modelu. Dzięki powtórzonemu wielokrotnie testowaniu modelu możliwe jest stwierdzenie jak bardzo powtarzalne są efekty treningu – pojawia się **informacja o rozrzucie uzyskiwanych wartości skuteczności**.

Dzięki takiemu podejściu łatwiej **porównywać między sobą klasyfikatory**, bo posiadanie jednocześnie miary np. średniej skuteczności modelu i informacji o wariancji skuteczności **umożliwia wykorzystanie testu statystycznego do stwierdzenia czy jeden z modeli radzi sobie istotnie statystycznie lepiej niż drugi.**

Test statystyczny polega na obliczeniu specjalnie przystosowanej do danego przypadku statystyki testowej i porównaniu jej z progiem (obliczanym na podstawie parametru α **- tzw. poziomu istotności**). Każdy test ma w **dwie wynikowe hipotezy** – **zerową** (H_0 - gdy statystyka jest mniejsza niż próg) i **alternatywną** (H_1 - gdy statystyka jest większa niż próg). Hipotezy mogą być różne w zależności od potrzeb (a tym samym – wybranego testu).

Testy statystyczne



Skuteczność klasyfikacji i macierze pomyłek

Najprostszą miarą tego, jak dobrze klasyfikator realizuje swoje zadanie jest **obliczenie np. procentowej poprawności (skuteczności) rozpoznawania** przez niego klas na podstawie dostarczonych danych walidacyjnych i testowych.

Klasyfikator operujący na zbiorze zawierającym taką samą liczbę przykładów dla każdej z klas zaczyna być skuteczny w momencie, gdy skuteczność jego klasyfikacji jest statystycznie **istotnie lepsza niż klasyfikacja przykładów poprzez losowe kwalifikowanie ich do jednej z możliwych klas.**

W udowodnieniu takiej tezy pomocne będą takie **techniki statystyczne jak np. przedziały ufności, test dwumianowy, czy test *t*-Studenta.** Służą one do określania na przykład, czy

- model **częściej** klasyfikuje dane poprawnie niż niepoprawnie (test **dwumianowy**),
- model A ma **wyższą średnią** dokładność klasyfikacji/regresji niż model B (test ***t*-Studenta)**.

Skuteczność klasyfikacji i macierze pomyłek

Jeżeli zbiór zawiera **więcej niż 2 klasy,** to często zamiast pojedynczą liczbą operujemy za pomocą tzw. **macierzy pomyłek** (ang. *confusion matrix*), która pozwala na określenie między innymi na to, które klasy są dla algorytmu problematyczne (np. najczęściej są ze sobą mylone). Przykład macierzy pomyłek dla algorytmu realizującego rozpoznawanie pojazdów na drodze:

		Efekt klasyfikacji		
		ciężarowy	**osobowy**	**motocykl**
Klasa wejściowa (oczekiwana)	**ciężarowy**	20	2	3
	osobowy	3	19	3
	motocykl	5	4	16

Macierz pomyłek to przykład tzw. tablicy kontyngencji i stąd w udowodnieniu **istotności statystycznej wyników** pomocne będą np. test McNemara, test χ^2, czy dokładny test Fishera.

Miary skuteczności dla przypadku, gdy klasy mają niezrównoważone liczności

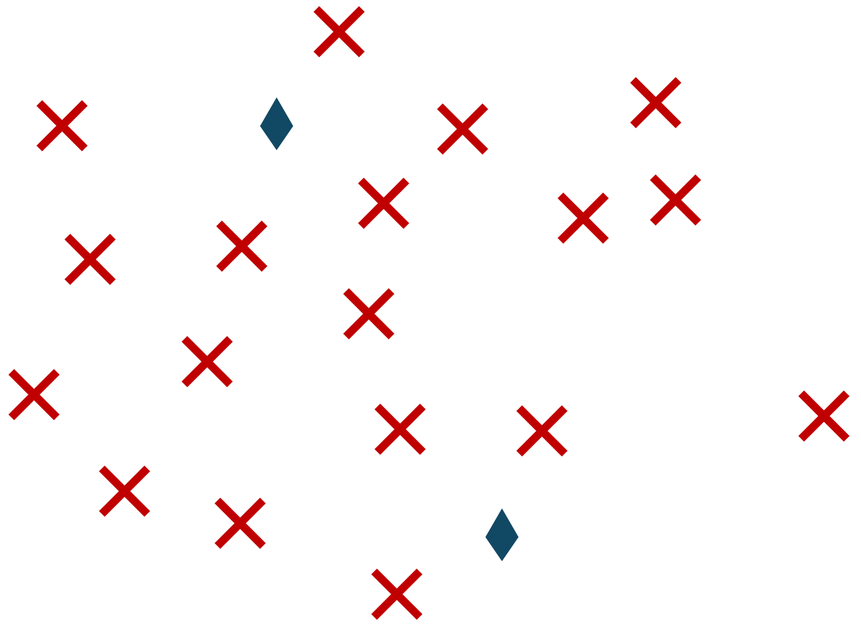
Czasami sama informacja o skuteczności klasyfikacji jest niewystarczająca. Może się tak zdarzyć nawet w przypadku, gdy skuteczność jest przedstawiona w postaci macierzy pomyłek.

Częstą sytuacją wymuszającą rezygnację z prostej miary, jaką jest skuteczność klasyfikacji, jest przypadek w którym realizowane jest zadanie klasyfikacji i **w wejściowym zbiorze danych liczebności przykładów powiązanych z poszczególnymi klasami bardzo od siebie odbiegają.**

Mówimy w takim przypadku, że mamy do czynienia ze **zbiorem danych o niezrównoważonej liczności klas** (ang. *imbalanced dataset*).

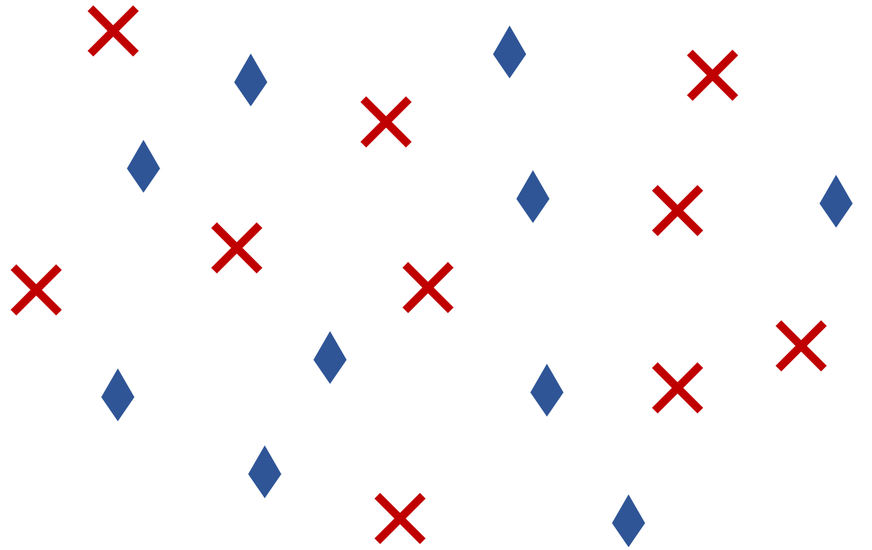
Oczywiście możliwe jest także **poradzenie sobie z tym problemem** poprzez przekształcenie zbioru danych, na przykład za pomocą **operacji upsamplingu i downsamplingu poszczególnych klas.**

Miary skuteczności dla przypadku gdy klasy mają niezrównoważone liczności



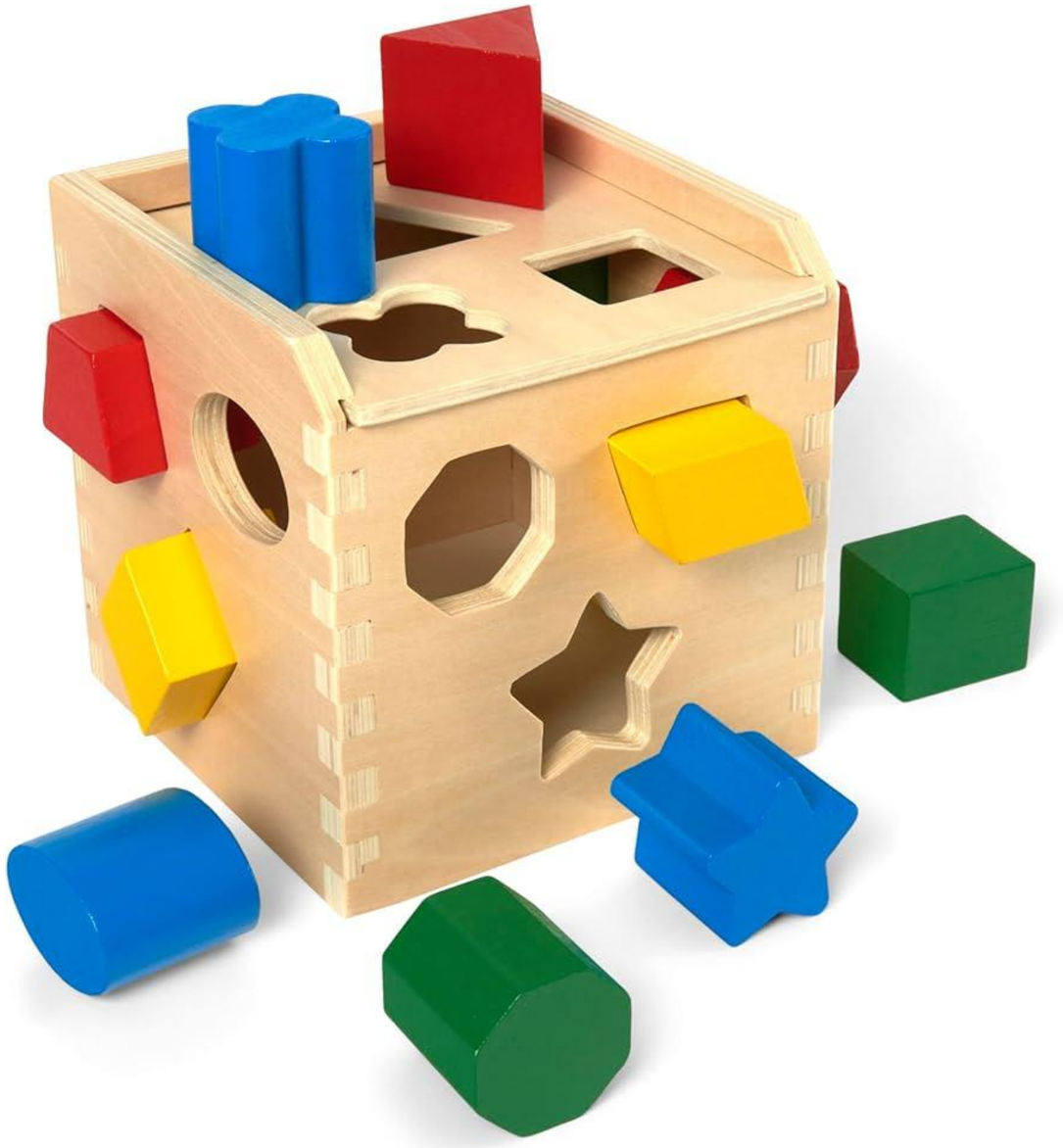
Jeśli zawsze zgaduję ✕, to mam 90% szans na to, że wybiorę poprawną odpowiedź.

Jeśli zawsze zgaduję ♦, to mam 10% szans na to, że wybiorę poprawną odpowiedź.



Jeśli zawsze zgaduję ✕, to mam 50% szans na to, że wybiorę poprawną odpowiedź.

Jeśli zawsze zgaduję ♦, to mam 50% szans na to, że wybiorę poprawną odpowiedź.



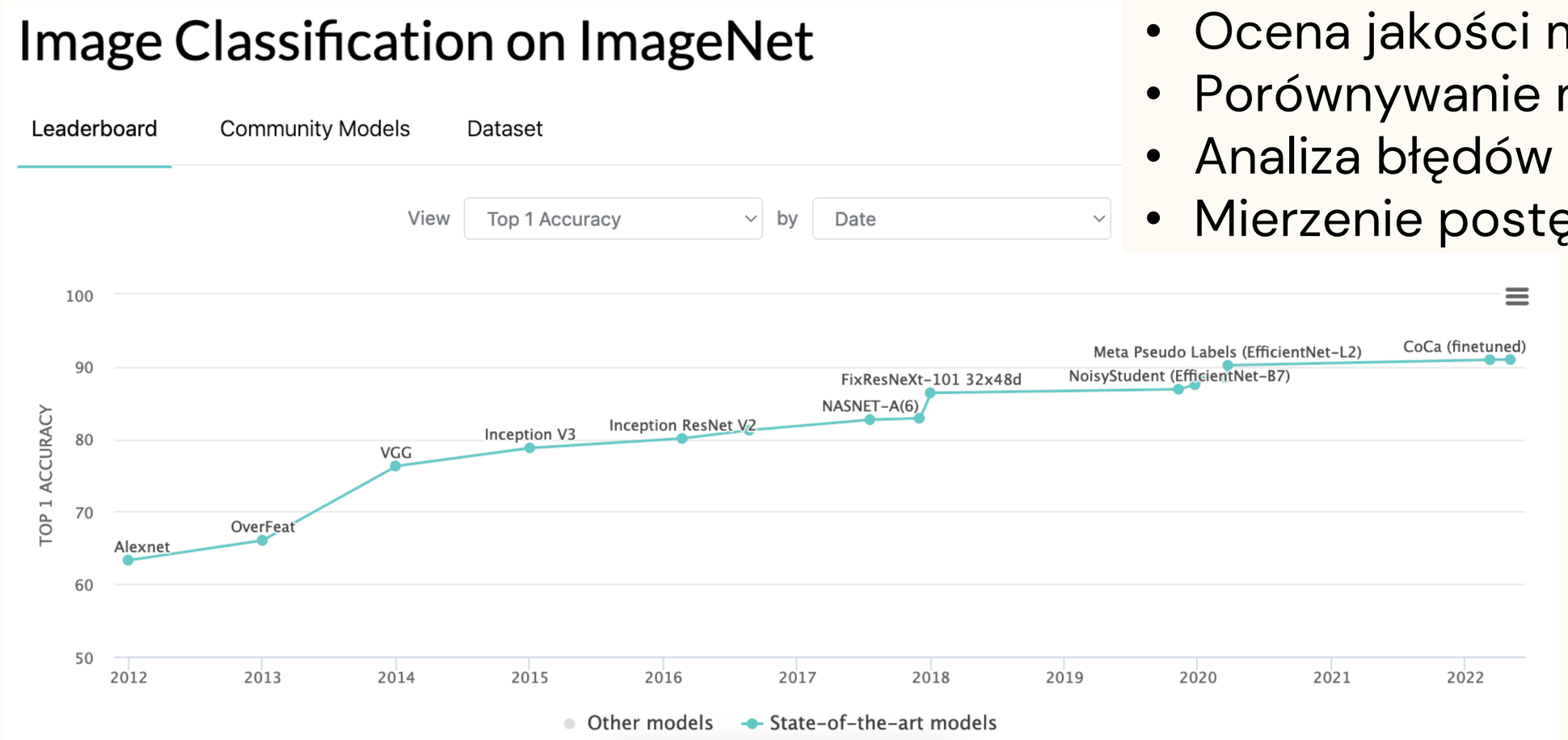
Klasyfikacja

- **Task**: przypisanie danych do wcześniej zdefiniowanych klas (kategorii) na podstawie określonych cech lub atrybutów.

- Celem klasyfikacji jest **znalezienie wzorców lub reguł, które umożliwią dokładne przewidywanie** przynależności nowych obiektów do określonych klas

Znaczenie i zastosowanie miar i metryk

Znaczenie i zastosowanie miar i metryk

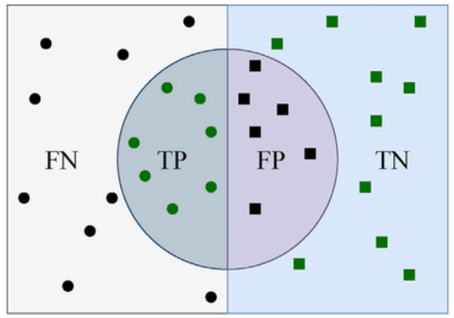


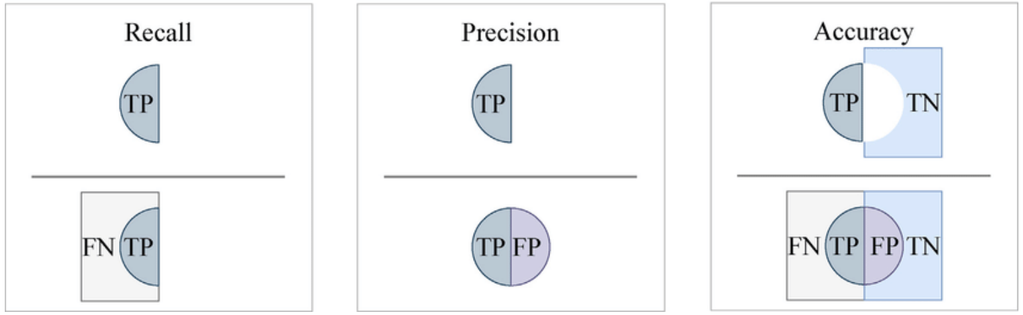
- Ocena jakości modelu
- Porównywanie modeli
- Analiza błędów
- Mierzenie postępów prac

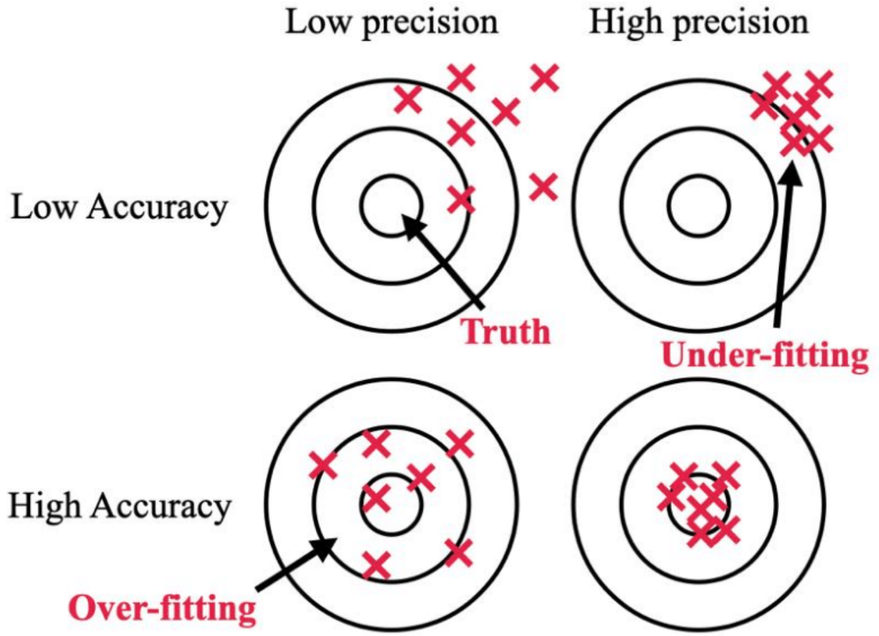
Miary dokładności dla klasyfikacji binarnej

Miara	Definicja	Wzór matematyczny	Przykładowa interpretacja
Dokładność (Accuracy)	Odsetek **poprawnie zaklasyfikowanych** obiektów	$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$	ogólna skuteczność klasyfikacji, w przypadku niezbalansowanych danych może być myląca
Precyzja (Precission)	Odsetek **poprawnie zaklasyfikowanych pozytywnych przypadków** w stosunku do wszystkich pozytywnych przypadków	$\text{Precission} = \frac{TP}{TP + \boldsymbol{FP}}$	szczególnie przydatna w przypadkach, gdy istnieją koszty związane z fałszywie pozytywnymi wynikami.
Czułość (Recall)	Odsetek poprawnie zaklasyfikowanych **pozytywnych** przypadków w stosunku do wszystkich rzeczywiście pozytywnych przypadków	$\text{Recall} = \frac{TP}{TP + \boldsymbol{FN}}$	szczególnie przydatna w przypadkach, gdy istnieją koszty związane z fałszywie negatywnymi wynikami.
Specyficzność (Specificity)	Odsetek poprawnie zaklasyfikowanych **negatywnych** przypadków w stosunku do wszystkich rzeczywiście negatywnych przypadków	$\text{Specificity} = \frac{TN}{TN + FP}$	
Wskaźnik F1 (F1 Score)	Średnia harmoniczna, pomaga zbalansować, precyzję i czułość.	$\text{F1} = \frac{2 * Precistion * Recall}{Precision + Recall}$	jest przydatna w przypadkach, gdy równie ważne jest zarówno unikanie fałszywie pozytywnych, jak i fałszywie negatywnych wyników

Miary dokładności dla klasyfikacji binarnej





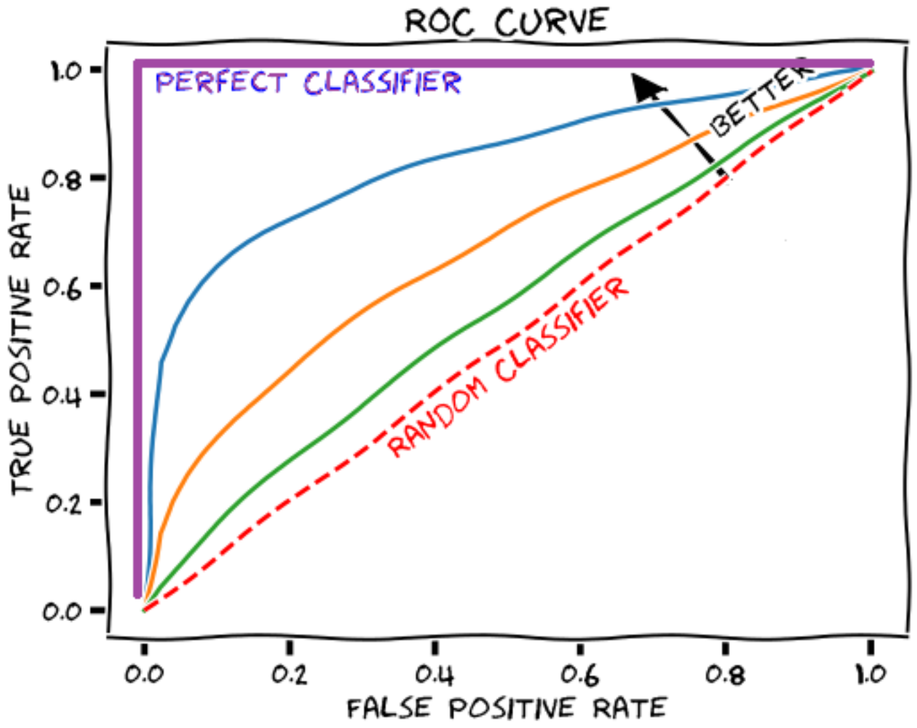


Miary dokładności dla klasyfikacji binarnej

AUC ROC – Area Under the Curve - Receiver Operating Characteristic

AUC ROC - jest to miara używana do oceny jakości modelu klasyfikacyjnego, zwłaszcza w problemach klasyfikacji binarnej.

Interpretacja:
•AUC = 1: Model idealnie rozróżnia klasy.
•AUC = 0.5: Model losowy, nie ma zdolności klasyfikacyjnych.
•AUC < 0.5: Model klasyfikuje gorzej niż losowy (co może sugerować odwrócenie prognoz).



Miary dla niezbalansowanych zbiorów

Miary istotne w przypadku niezbalansowanych klas, gdzie występuje duża różnica w liczbie próbek między klasami.

Miara	**Definicja**	**Wzór matematyczny**
Współczynnik Youdena	Miara równoważąca czułość i specyficzność	czułość + specyficzność − 1
Precyzja ważona	Miara precyzji, która uwzględnia nierównomierność klas	(liczba próbek klasy 1 ∗ precyzja 1 + liczba próbek klasy 0 ∗ precyzja 0) / (liczba próbek klasy 1 + liczba próbek klasy 0)
Miara przestrzeni kosztowej	Miara uwzględniająca koszty błędów klasyfikacji	(koszt fałszywie pozytywnej ∗ liczba FP + koszt fałszywie negatywnej ∗ liczba FN) / (liczba próbek)

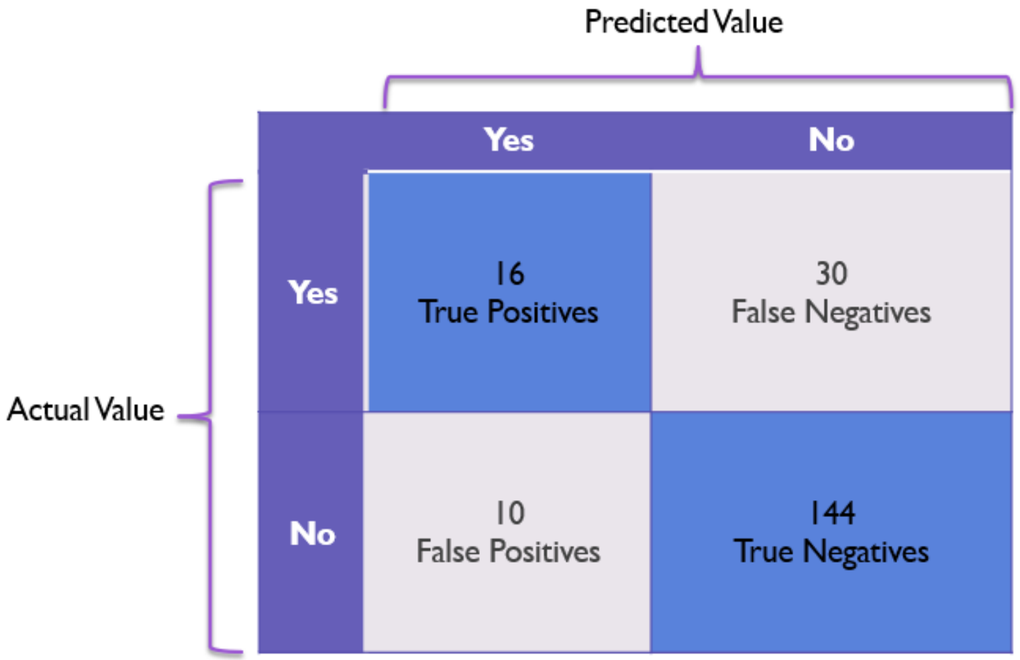
Metryki wieloklasowe

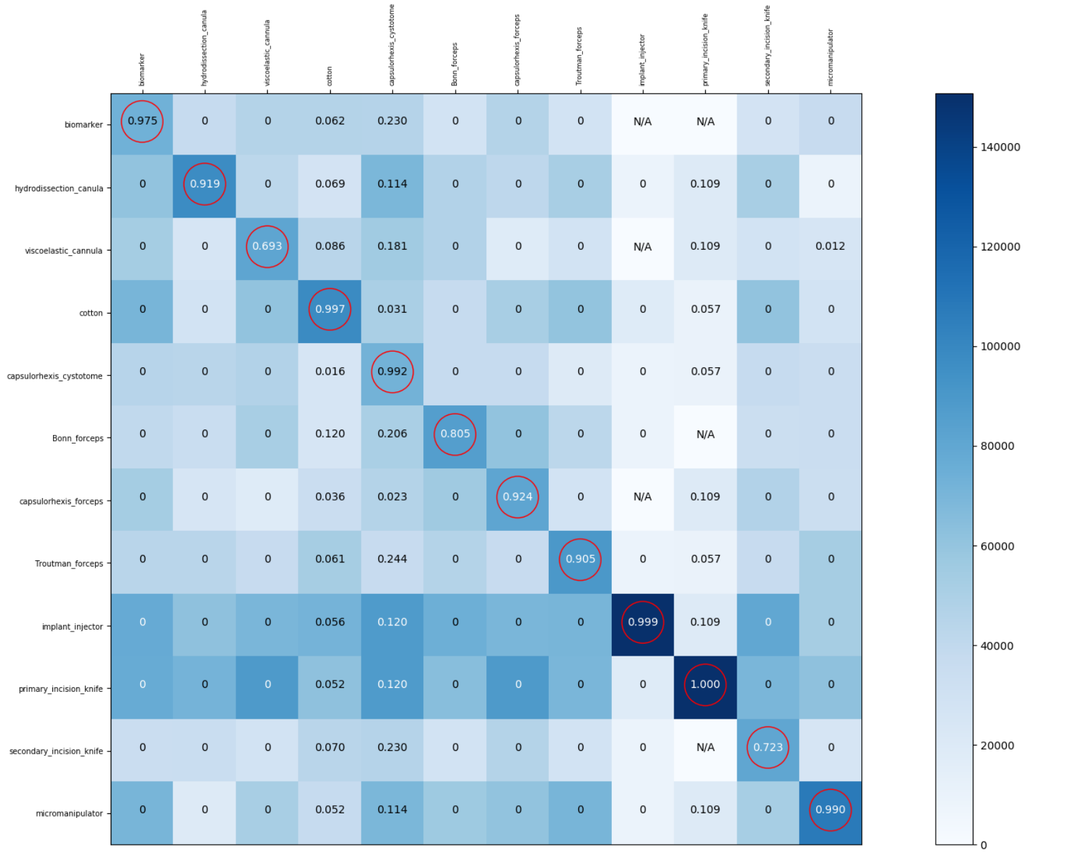
• Te metryki są przydatne w klasyfikacji wieloklasowej, gdzie mamy więcej niż dwie klasy i chcemy ocenić jakość modelu oraz stopień zgodności między przewidywanymi etykietami a rzeczywistymi etykietami.

Miara	**Definicja**
Macierz pomyłek	Miara równoważąca czułość i specyficzność
Wskaźnik Jaccarda	Miara podobieństwa między dwoma zbioram
Miara Hamminga	Liczba niezgodnych elementów między dwoma ciągami o jednakowej długości

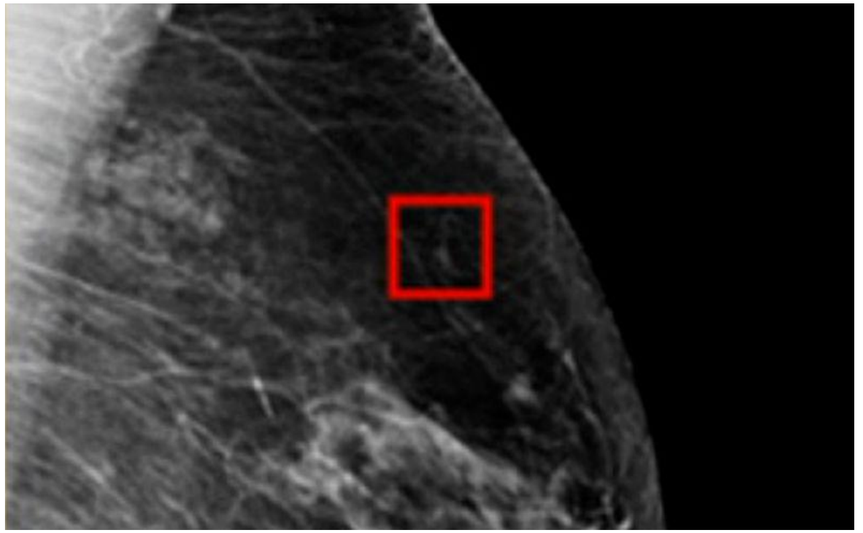
Macierz pomyłek

- Macierz pomyłek jest używana do oceny jakości klasyfikacji, a na jej podstawie można obliczyć różne metryki, takie jak dokładność, precyzja, czułość, specyficzność i wiele innych.

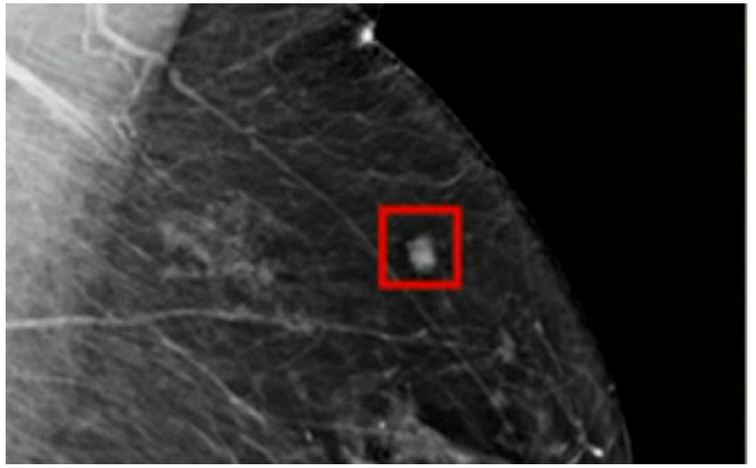


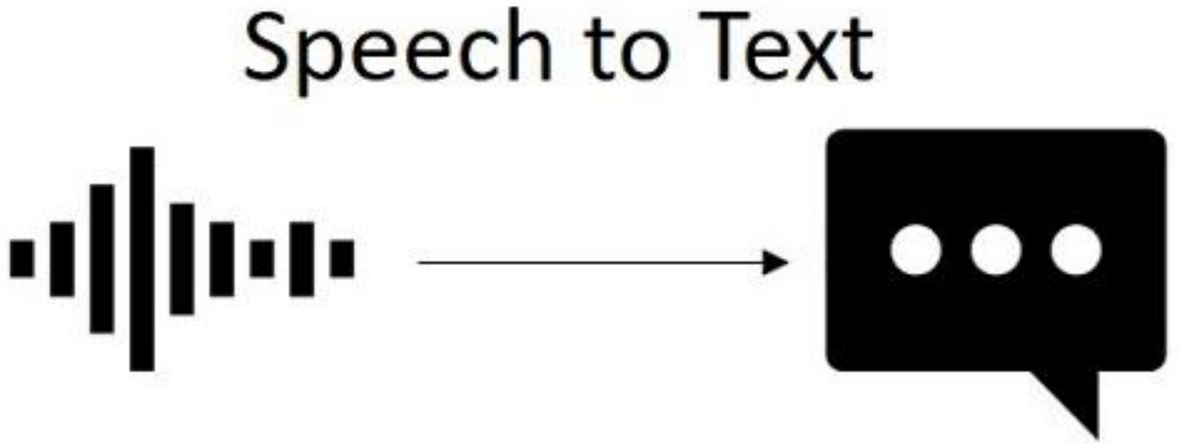


Wybór odpowiednich miar i metryk



- Wybór miar zależy od kontekstu problemu:
  - W **medycynie** istotne może być znalezienie miary o wysokiej czułości, aby jak najmniej pomijać przypadki pozytywne, nawet kosztem większej liczby fałszywie pozytywnych wyników
  - W analizie danych finansowych ważniejsza może być wysoka precyzja w celu uniknięcia błędów kosztownych dla firmy.





Transkrypcja mowy

Task: przekształcenie **mowy** na **tekst** przy użyciu algorytmów AI i/lub uczenia maszynowego

Miary dokładności dla transkrypcji mowy

$$WER = \frac{S_w + D_w + I_w}{N_w}$$

$$CER = \frac{S_c + D_c + I_c}{N_c}$$

S = Number of Substitutions

D = Number of Deletions

I = Number of Insertions

N = Number of characters in reference text

Example:

Reference: “my name is kenneth”

Observation: “myy nime iz kenneth”

WER = 75%

CER = 16.67%

ASR for Polish speech

Experiment / Used datasets	Testest [WER]			
	torgo	UASpeech	pl_dysarthric_v1	pl_dysarthric_synth_test_v1
vanilla-whisper-tiny	100.15	211.01	179.82	161.31
torgo_UASpeech	36.40	44.37	97.51	108.56
torgo_UASpeech_classicsynth	34.50	35.10	112.77	118.98
dysarthric_synth_dataset_1k+trogo+UASpeech	**25.15**	**29.90**	91.34	40.71
dysarthric_synth_dataset_10k+trogo+UASpeech	29.53	31.11	**82.91**	**16.24**

Overfitting?

LOWER IS BETTER

Podsumowanie

Miary klasyfikacji są kluczowe do analizy, a co za tym idzie doboru modelu i jego parametrów do danych

Aby poprawnie korzystać z tych miar, należy dobrze rozumieć dane i ich kontekst