WSI lab 4

Marcin Jarczewski

> ***Dlaczego na jednym zbiorze jest znacznie lepszy wynik niż na drugim?***
> *Do potwierdzenia lub odrzucenia postawionych hipotez konieczne może być przeprowadzenie dodatkowych eksperymentów ze zmodyfikowanymi zbiorami danych. Sformułować i spisać wnioski. Atrybuty nominalne, testy tożsamościowe. Podać dokładność i macierz pomyłek na zbiorach: Breast cancer i mushroom*

Analiza całych zbiorów danych

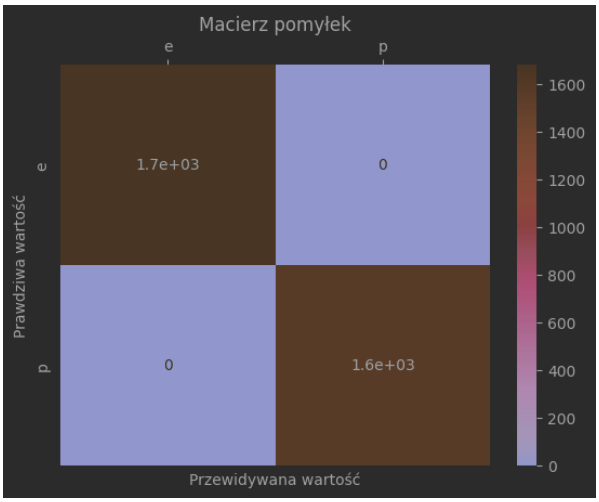
Zbiór danych: agaricus-lepiota	Liczność klasy	% wszystkich przykładów
e	4208	51,8
p	3916	48,2

Zbiór danych: breast-cancer	Liczność klasy	% wszystkich przykładów
no-recurrence-events	201	70,28
recurrence-events	85	29,72

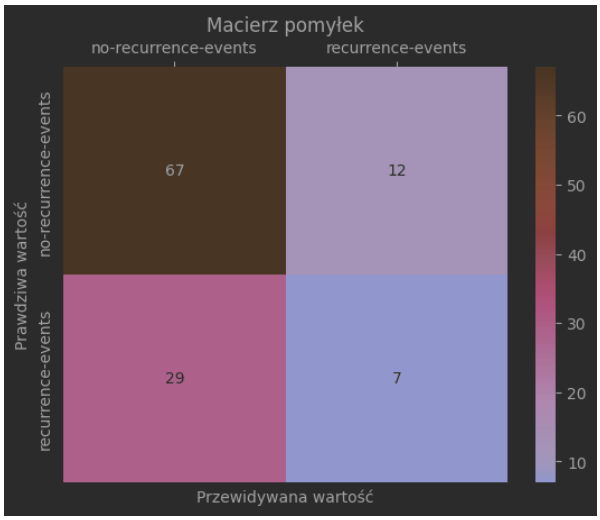
Eksperyment (Dane nie zmodyfikowane)

Dla nie zmodyfikowanych danych, klasyfikator ID3 na podzbiorze testowym, który został podzielony w stosunku 3 do 2 osiągnął następującą dokładność:

Zbiór danych	Dokładność [%]	Liczba przykładów trenujących	Liczba przykładów testowych	Stosunek zbioru trenującego do testowanego
agaricus-lepiota	100,00	4874	3250	1,500
breast-cancer	64,35	171	115	1,487



Dla zbioru agaricus-lepiota



Dla zbioru breast-cancer

Analiza zbiorów testujących

Nazwa klasy	Liczność	%
e	1678	51,63
p	1572	48,37

Nazwa klasy	Liczność	%
no-recurrence-events	79	68,70
recurrence-events	36	31,30

Komentarz:

Osiągnięty wynik dla pierwszego zbioru jest zaskakujący. Można stwierdzić że jest to aż podejrzane. Stuprocentowa dokładność może świadczyć o zjawisku przeuczenia się (overfitting). Zbiór danych testowych został równo podzielony względem zawartości klas (obie klasy ~50%)

Natomiast dla drugiego zbioru wynik wydaje się sensowny. Pozyskane informacje sprawiają że dokładność modelu jest większa niż gdybyśmy tylko zgadywali. Na mniejszą dokładność może wpływać mniejsza objętość zbioru trenującego oraz większa liczność jednej klasy zarówno w obu zbiorach.

Eksperyment (Breast cancer test - zmodyfikowane dane po równo)

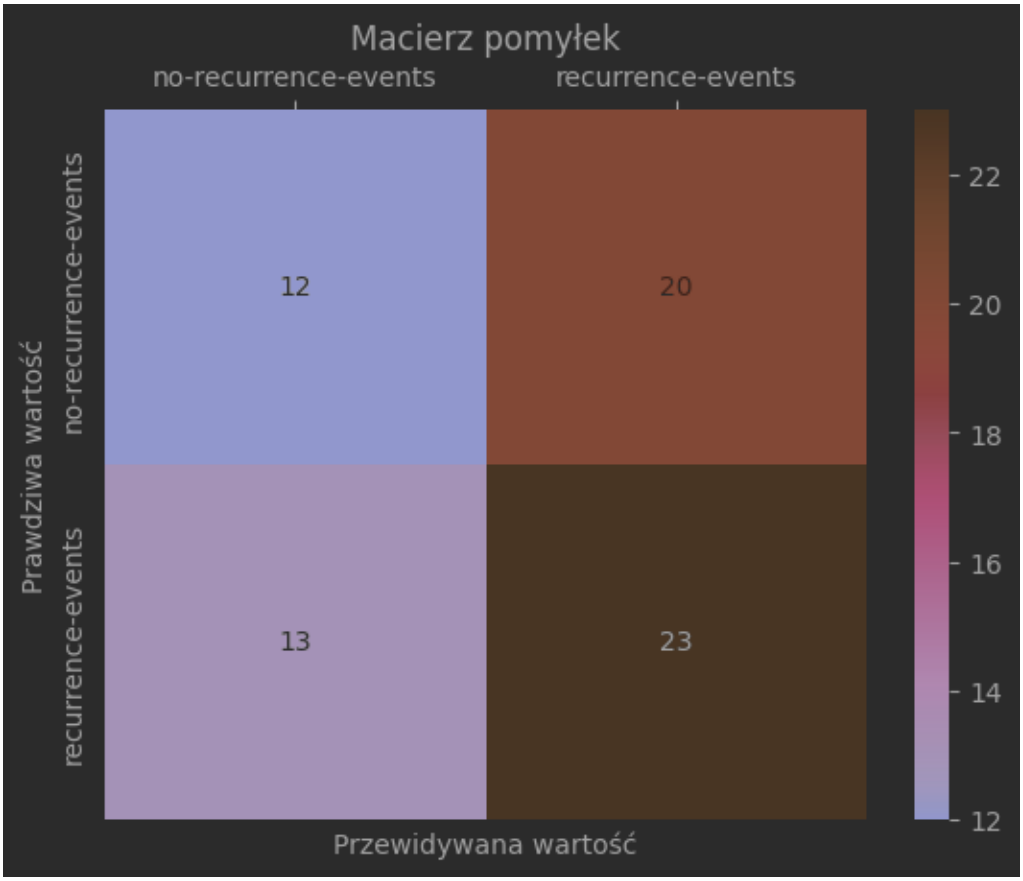
Zbiór danych	Dokładność [%]	Liczba przykładów trenujących	Liczba przykładów testowych	Stosunek zbioru trenującego do testowanego
breast-cancer_test	51,47	102	68	1,5

Zbiór danych został zmodyfikowany w taki sposób aby uzyskać taką samą liczność klas:

Analiza zbioru danych

Zbiór danych: breast-cancer_test	Liczność klasy	% wszystkich przykładów

Zbiór danych: breast-cancer_test	Liczność klasy	% wszystkich przykładów
no-recurrence-events	85	50
recurrence-events	85	50



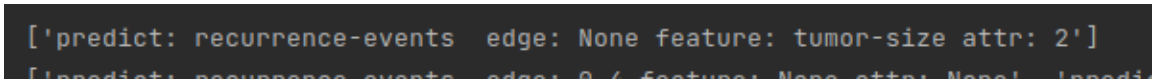
Analiza zbioru testującego

Nazwa klasy	Liczność	% wszystkich przykładów
no-recurrence-events	32	47,06
recurrence-events	36	52,94

Komentarz:

Dla równego podziału klas nie osiągnięto lepszych wyników. Jest to spowodowane faktem, że algorytm ID3 jest algorytmem zachłannym - preferuje podział względem atrybutów, które najbardziej wpływają na entropię zbioru. Stąd podział losowy względem samych klas nie

spowoduje lepszych wyników. Ponadto należy wspomnieć że dla podanego zbioru najbardziej znaczącym atrybutem jest *tumor-size.*



początkowy fragment drzewa decyzyjnego - na zrzucie ekranu korzeń

Eksperyment (Breast cancer test 2 - zmodyfikowane dane)

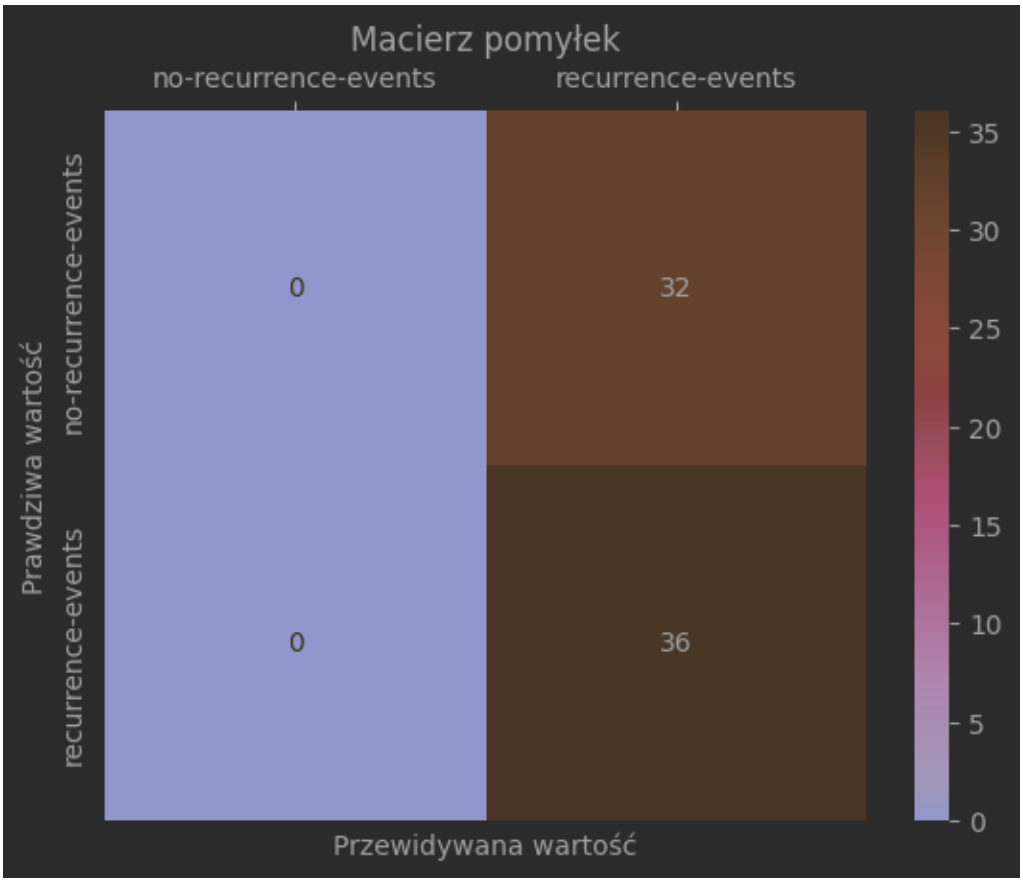
Ze zbioru danych podzielonego na pół (*breast_cancer_test*) została usunięta kolumna *tumor-size*

Zbiór danych	Dokładność [%]	Liczba przykładów trenujących	Liczba przykładów testowych	Stosunek zbioru trenującego do testowanego
breast-cancer_test2	52.94	102	68	1,5

Zbiór danych został zmodyfikowany w taki sposób aby uzyskać taką samą liczność klas:

Analiza zbioru danych

Zbiór danych: breast-cancer_test2	Liczność klasy	% wszystkich przykładów
no-recurrence-events	85	50
recurrence-events	85	50



Analiza zbioru testującego

Nazwa klasy	Liczność	% wszystkich przykładów
no-recurrence-events	32	47,06
recurrence-events	36	52,94

Komentarz:

W tym przypadku widać że algorytm przewiduje stale jedną wartość. W ten sposób potwierdza się zachłanność algorytmu oraz wpływ kolumny *tumor_size* która sprawiała że uzyskiwane wyniki nie były jednomyślne (każdy kwadrat, miał zbliżoną wartość w poprzednim przypadku)

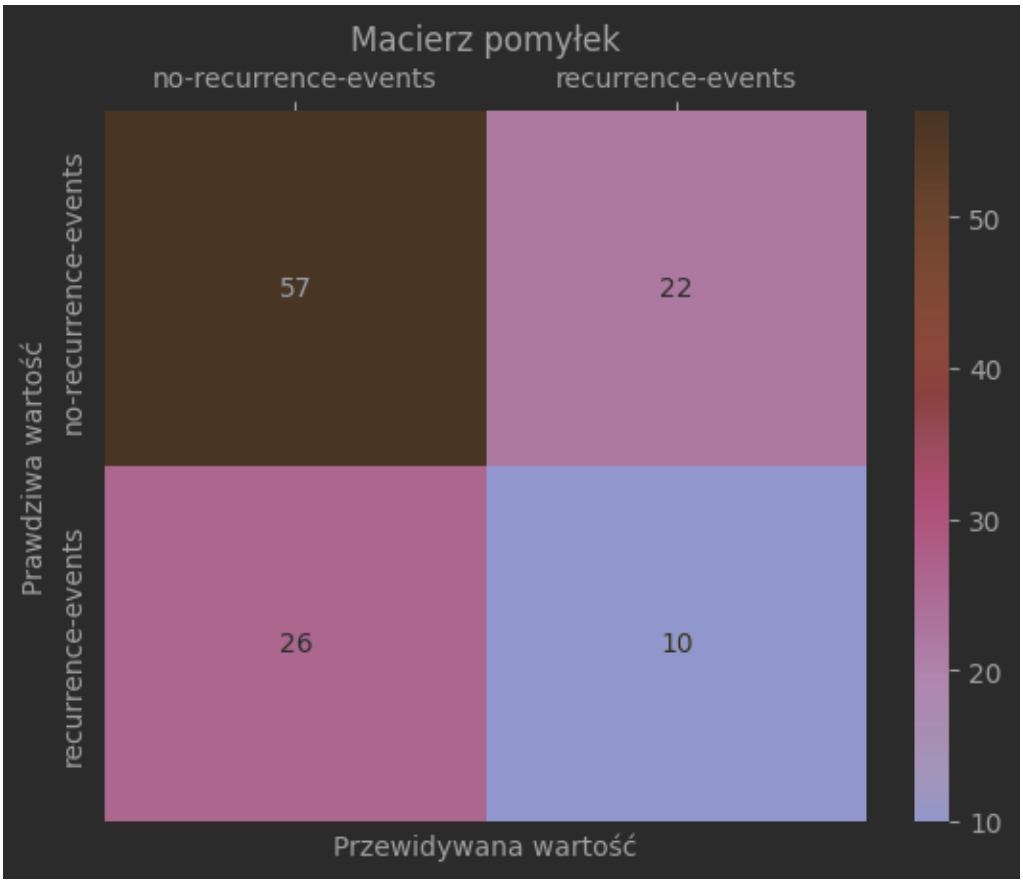
Eksperyment (Breast cancer test 3 - zmodyfikowane dane)

Ze zbioru danych (*breast_cancer*) została usunięta kolumna *inv-nodes* (najbardziej znacząca)

Zbiór danych	Dokładność [%]	Liczba przykładów trenujących	Liczba przykładów testowych	Stosunek zbioru trenującego do testowanego
bc_without_best_column	58.26	171	115	1,487

Analiza zbioru danych

Zbiór danych: bc_without_best_column	Liczność klasy	% wszystkich przykładów
no-recurrence-events	201	70,28
recurrence-events	85	29,72



Analiza zbioru testującego

Nazwa klasy	Liczność	% wszystkich przykładów

Nazwa klasy	Liczność	% wszystkich przykładów
no-recurrence-events	79	68,7
recurrence-events	36	31,3

Komentarz:

Usunięcie głównej kolumny ze zbioru wpływa na “przesunięcie się” wartość z poprawnie ocenionych (z lewego górnego rogu na prawy górny róg, True Positive →False Negative). Za każdym razem dokładności przewidywań przyjmują podobne wartości ~50 kilka procent.

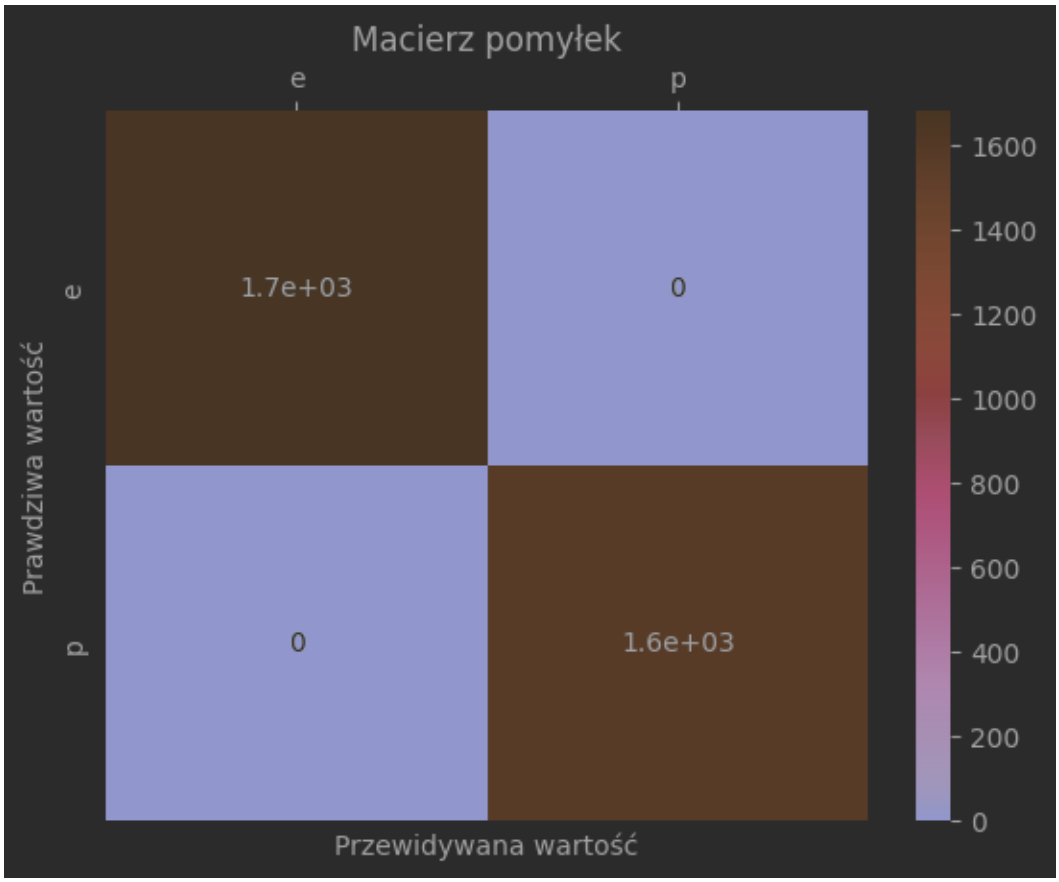
Eksperyment (Dane nie zmodyfikowane)

Eksperyment (agaricus-lepiota - zmodyfikowane dane)

Zbiór danych	Dokładność [%]	Liczba przykładów trenujących	Liczba przykładów testowych	Stosunek zbioru trenującego do testowanego
Out_19	100,0	4874	3250	1,5

Analiza zbioru danych

Zbiór danych: agaricus-lepiota	Liczność klasy	% wszystkich przykładów
e	4208	51,8
p	3916	48,2



Analiza zbioru testującego

Nazwa klasy	Liczność	% wszystkich przykładów
e	1678	51,63
p	1572	48,37

Komentarz:

Pomimo usunięcia głównej kolumny ze zbioru danych dokładność jest dalej 100% na co wpływa duża próba oraz duża liczba kolumn, przez co pojedyncza kolumna nie ma tak dużego znaczenia co w przypadku drugiego zbioru. Świadczy to o wysokiej “stabilności danych”.

Wnioski końcowe

Badając dokładność przewidywań, klasyfikator ID3 osiąga na zbiorze dotyczącym grzybów lepsze wyniki z następujących przyczyn:

1. Liczba przypadków testowych im większa tym lepszą próbę możemy przyjąć. W małolicznym zbiorze można nie zaobserwować ogólnych trendów
2. Inna liczba atrybutów sprawia, że “ważności” kolumn są inne przez co nie można dokonać aż tak dokładnego podziału