

The background of the slide features a light beige, textured paper-like surface. In the upper right, there is a faint, dark silhouette of a mountain range. On the right side, a dark, thin branch of a willow tree hangs down, adorned with small, dark, round buds.

Drzewa Decyzyjne, cz.2

Inteligentne Systemy Decyzyjne

Katedra Systemów Multimedialnych

WETI, PG

Opracowanie: dr inż. Piotr Szczuko

Podsumowanie poprzedniego wykładu

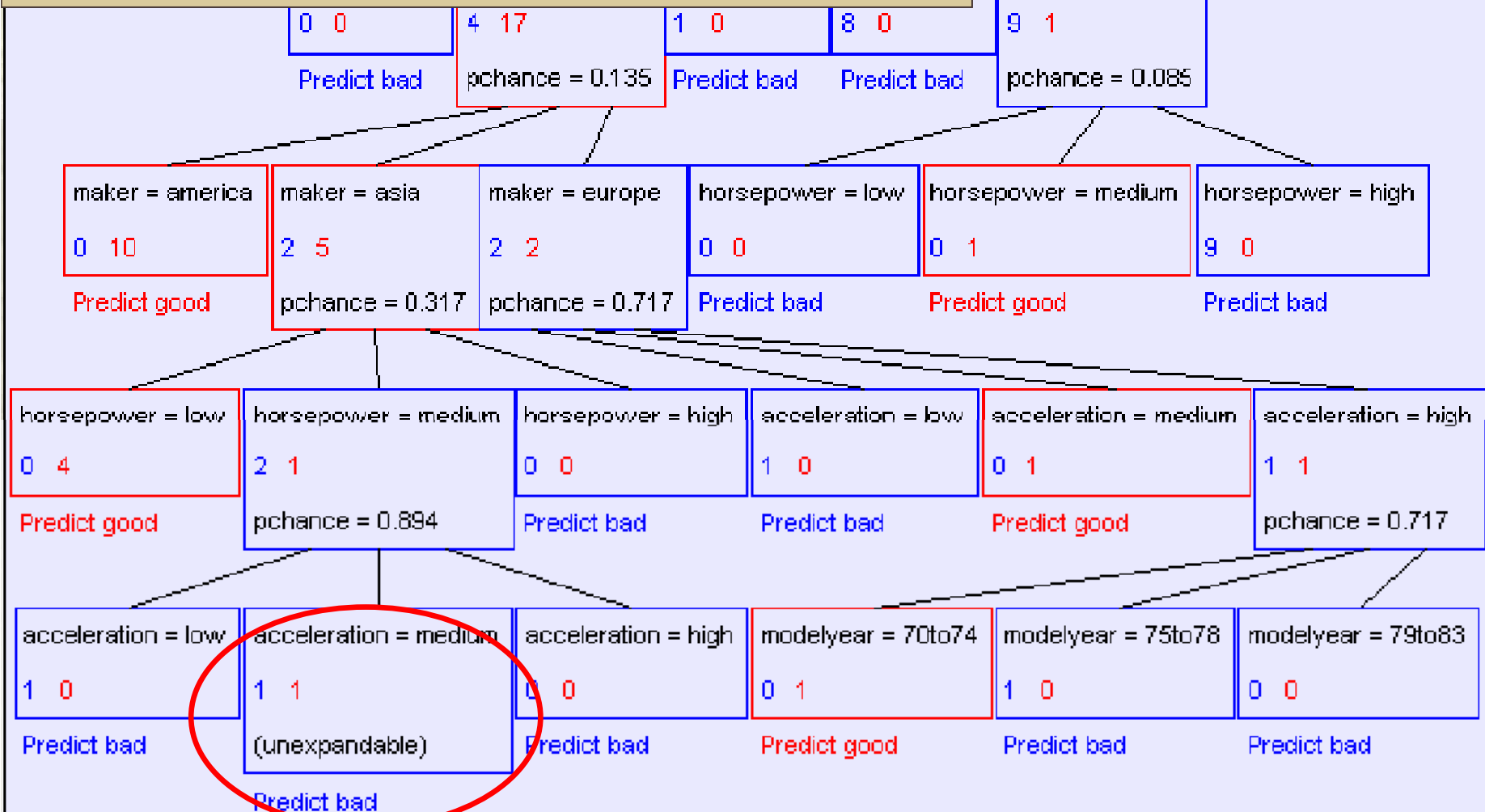
- ❖ Cel: przewidywanie wyniku (określania **kategorii**, klasyfikowanie) na podstawie posiadanych parametrów opisujących obiekt
- ❖ Analiza tablic kontyngencji
- ❖ Duża i mała entropia
- ❖ Zysk informacyjny $IG(Y|X)$
- ❖ Budowanie drzewa decyzyjnego
- ❖ Błąd treningowy i testowy

mpg values: bad good

Gotowe Drzewo Decyzyjne

root
22 18
pchance = 0.001

Liczba błędów 1
Liczba obiektów 40
Procent błędnych decyzji 2,5% (zbiór treningowy)



Gotowe Drzewo Decyzyjne

mpg values: bad good

root
22 18
pchance = 0.0001

Liczba błędów	Liczba obiektów	Procent błędnych decyzji
1	40	2,5% (zbiór treningowy)
74	352	21,02% (zbiór testowy)

3 cylinders = 8

Procent błędnych decyzji
2,5% (zbiór treningowy)
21,02% (zbiór testowy)

mak
0 1
Pred

Skąd wynika tak duża różnica?

Czy można poprawić efektywność klasyfikacji?

Czy drzewo może być mniejsze?

horsepow
0 4
Predict g

igh
7

accelera
1 0

83

Predict bad

(unexpandable)

Predict bad

Predict good

Predict bad

Predict bad

Predict bad

Przykład – zbiór treningowy

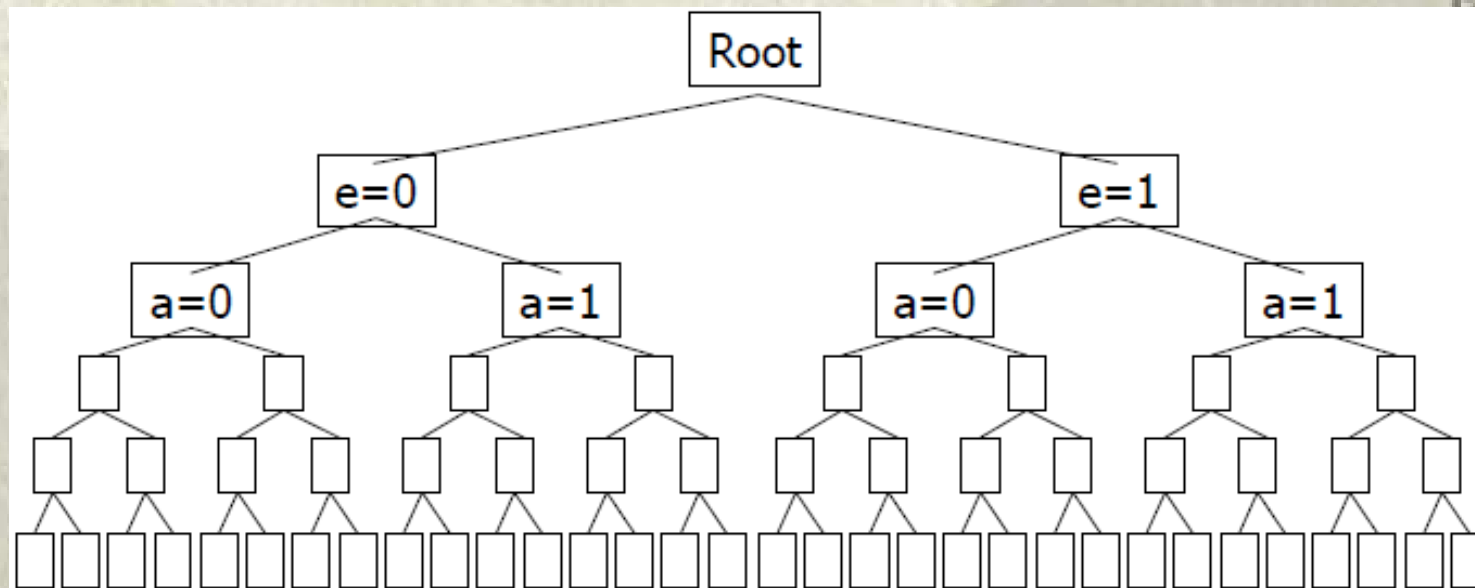
- ❖ Zbiór danych treningowych wytworzony w kontrolowany sposób:
 - wszystkie kombinacje 5 bitów **abcde**
 - wyjście **y** jako kopia **e**, poza 25% przypadków, gdzie zastosowana jest inwersja logiczna **!e**

a	b	c	d	e	y
0	0	0	0	0	0
0	0	0	0	1	0
0	0	0	1	0	0
0	0	0	1	1	1
0	0	1	0	0	1
:	:	:	:	:	:
1	1	1	1	1	1

Przykład – zbiór testowy

- ❖ Zbiór testowy:
 - wszystkie kombinacje 5 bitów **abcde**
 - wyjście **y** jako kopia **e**, poza 25% przypadków (innych niż wcześniej), gdzie zastosowana jest inwersja logiczna **!e**
- ❖ Zbiory są prawie identyczne:
 - niektóre **y** „uszkodzone” w zbiorze treningowym nie będą „uszkodzone” w testowym (i odwrotnie)

Przykład – drzewo decyzyjne



– (Dlaczego w pierwszym rozgałęzieniu jest sprawdzany bit **e**?)

❖ Błąd treningowy wynosi 0%!

– Drzewo uwzględnia całość zbioru danych, wszystkie 32 kombinacje i właściwe wyjścia **y** dla nich

Przykład – błąd testowy

	$\frac{1}{4}$ liści „uszkodzonych”	$\frac{3}{4}$ liści dobrych
$\frac{1}{4}$ danych „uszkodzonych”	$\frac{1}{16}$ zbioru testowego zostanie przypadkowo dobrze sklasyfikowana	$\frac{3}{16}$ zbioru testowego zostanie błędnie sklasyfikowanych ponieważ dane są „uszkodzone”
$\frac{3}{4}$ danych dobrych	$\frac{3}{16}$ zbioru testowego zostanie błędnie sklasyfikowanych ponieważ liście są „uszkodzone”	$\frac{9}{16}$ zbioru testowego zostanie przypadkowo dobrze sklasyfikowanych

Spodziewamy się popełnić błędy w $\frac{3}{8}$ przypadków (37,5%)

Wnioski

- ❖ Zbiór treningowy i testowy – bardzo znacząca rozbieżność wyników
- ❖ Należy odpowiednio przygotować się do klasyfikacji danych w przyszłości
- ❖ - *jak?*

Przykład – redukcja danych

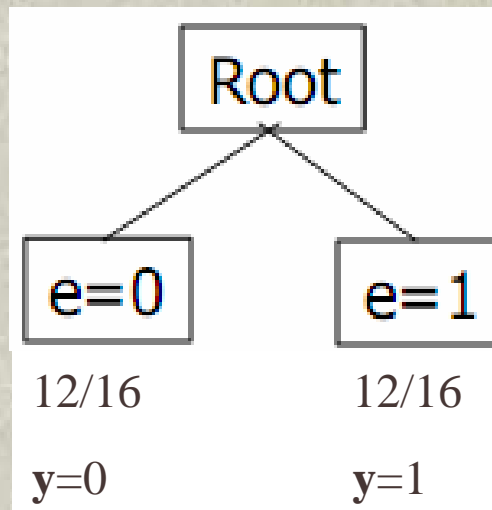
- ❖ Niech zbiór danych będzie następujący:

a	b	c	d	e	y
0	0	0	0	0	0
0	0	0	0	1	0
0	0	0	1	0	0
0	0	0	1	1	1
0	0	1	0	0	1
:	:	:	:	:	:
1	1	1	1	1	1

- Bity **a-d** ukryte
- Wyjście **y** równe bitowi **e** z wyjątkiem 25% przypadków

Drzewo decyzyjne

- ❖ Uwzględnia tylko te dane do których mamy dostęp:



- Nie pozwala na uwzględnienie „uszkodzonych” wyjściowych $y \neq e$
- Już na starcie klasyfikacja obciążona jest *błędem treningowym*

Błąd testowy

- ❖ Ten sam zbiór co poprzednio – 25% y jest negacją e
- ❖ Te 25% zostanie sklasyfikowanych błędnie

25% w porównaniu do 37,5% wcześniej

Przetrenowanie

- ❖ Jeżeli inteligentny system decyzyjny analizuje dane nieistotne (szum) wówczas zachodzi **przetrenowanie** (ang. *overfitting*)
- ❖ **Przetrenowany** system decyzyjny osiąga:
 - *wysoką* trafność klasyfikacji danych treningowych
 - *niską* trafność klasyfikacji danych testowych

Przetrenowanie

- ❖ Zwykle brak jest informacji ujawniających, które atrybuty są nieistotne
- ❖ Ponadto istotność zależności może od kontekstu, np.:
 - $y = a \text{ AND } b$
 - dla $a = 0$ informacja o wartości b jest nieistotna
 - dla $a = 1$ informacja o wartości b jest istotna

Przetrenowanie

- ❖ **Statystyka** może dostarczyć informacji o tym, które atrybuty są nieistotne
- ❖ Test χ^2 Pearsona, test istotności dla zmiennych jakościowych (skategoryzowanych).
- ❖ Miara ta oparta jest na możliwości obliczenia licznosci **oczekiwanych**,
 - tj. licznosci, jakich oczekiwalibyśmy, gdyby **nie istniała zależność** między zmiennymi.

Test χ^2 Pearsona

- ❖ Przypuśćmy, że pytamy 20 mężczyzn i 20 kobiet o upodobanie do jednej z dwóch gatunków wody mineralnej (gatunki A i B).
- ❖ Gdyby nie było **żadnej** zależności między upodobaniem odnośnie wody mineralnej a płcią, wówczas należałoby **oczekiwać** mniej więcej **jednakowych licznosci** w preferencjach gatunku A i B dla obu płci.
- ❖ Test Chi-kwadrat staje się istotny w miarę **wzrostu odstępstwa** od tego oczekiwanego schematu (to znaczy w miarę jak licznosci odpowiedzi dla mężczyzn i kobiet zaczynają się różnić).

Test χ^2 Pearsona

- ❖ Sprawdzana jest hipoteza zerowa o niezależności cech
- ❖ $n > 30$, n -elementowa próba z populacji
- ❖ Dwie cechy, indeksowane po i oraz po j

$$\chi^2 = \sum_{j=1}^k \sum_{i=1}^r \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}} = \sum_{j=1}^k \sum_{i=1}^r \left(\frac{n_{ij}^2}{\hat{n}_{ij}} - n_{ij} \right)$$

- ❖ n_{ij} – liczba elementów opisanych wartościami i, j kryteriów
- ❖ \hat{n}_{ij} – teoretyczna liczebność, wg. wzoru:

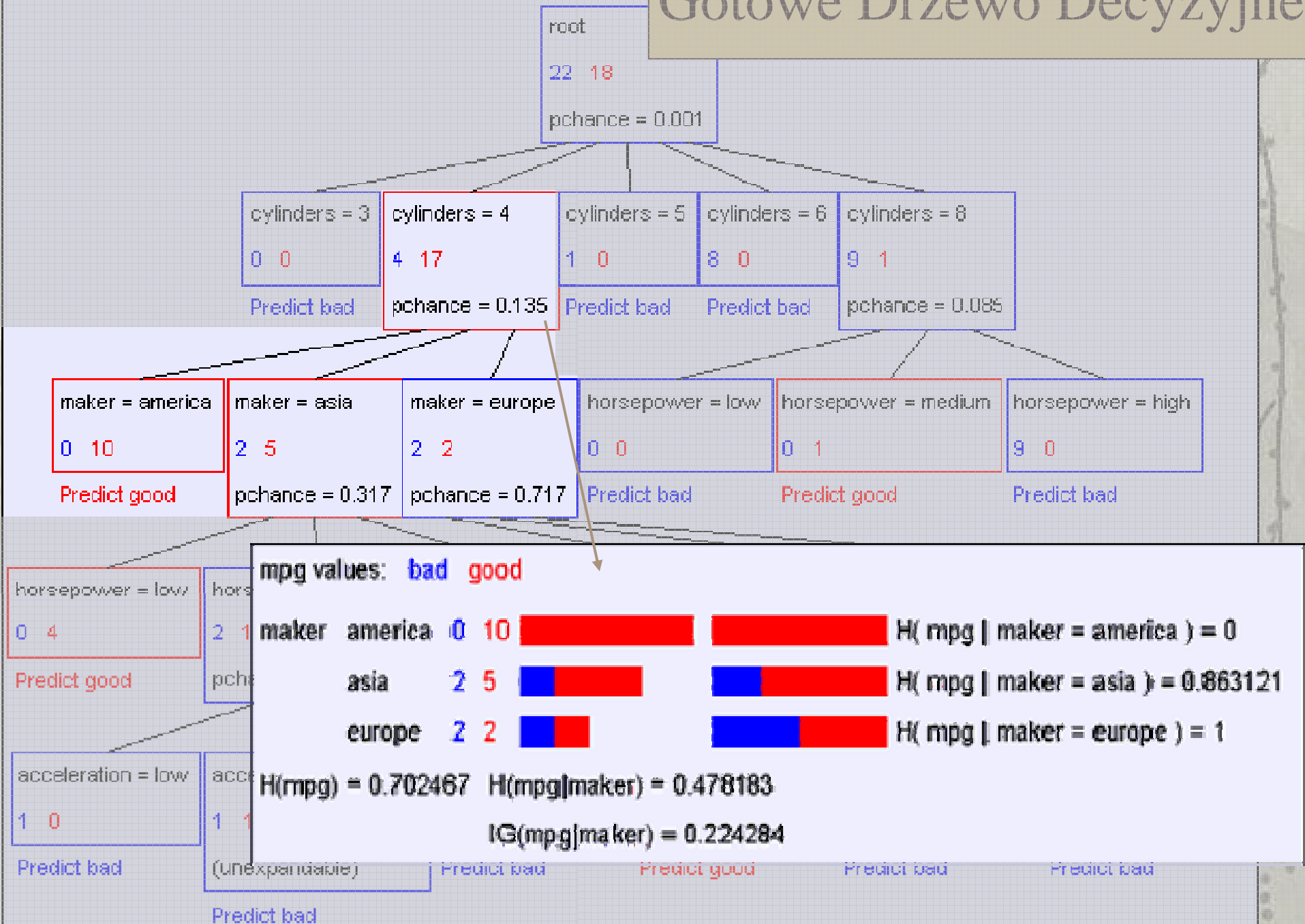
$$\hat{n}_{ij} = \frac{\sum_{j=1}^k n_{ij} \sum_{i=1}^r n_{ij}}{n}$$

Test χ^2 Pearsona

- ❖ Wartość χ^2 porównać należy z $\chi^2_{\alpha; (r-1)(k-1)}$ odczytaną z tablic statystycznych:
 - α to zakładany poziom istotności (np. 0,005; 0,01; 0,05)
 - $(r-1)(k-1)$ to liczba stopni swobody
- ❖ Jeżeli $\chi^2 \geq \chi^2_{\alpha; (r-1)(k-1)}$ to odrzucamy hipotezę H_0 o niezależności cech (**cechy są zależne**)
- ❖ Jeżeli $\chi^2 < \chi^2_{\alpha; (r-1)(k-1)}$ to nie ma podstaw do odrzucenia H_0

Gotowe Drzewo Decyzyjne

mpg values: bad good



Test Chi²

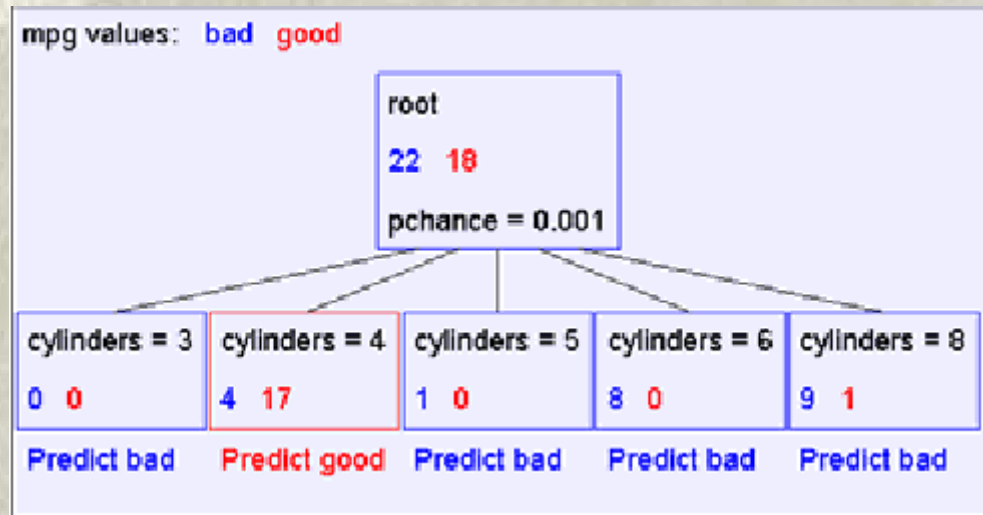
- ❖ Przypuśćmy, że MPG jest całkowicie niezależne (nieskorelowane) z producentem
- ❖ Jakie jest wówczas prawdopodobieństwo zaobserwowania takich danych? (danych, które są dziełem przypadku, a nie wynikają z zależności między atrybutami)
- ❖ $p = 13,5\%$

Wykorzystanie testu Chi^2

- ❖ Zbudować „pełne” drzewo decyzyjne
- ❖ Upraszczenie (ang. *Prunning*) drzewa:
 - Usuwać od dołu te rozgałęzienia, w których $p > \text{Max}P$
- ❖ Parametr *MaxP* dobrany w zależności od chęci podejmowania ryzyka dopasowania drzewa do szumu (danych nieistotnych)

Przykład

❖ Dla $MaxP = 0,1$ uzyskuje się drzewo:



Poprzednio:

Liczba błędów	Liczba obiektów	Procent błędnych decyzji
1	40	2,5% (zbiór treningowy)
74	352	21,02% (zbiór testowy)

Teraz:

Liczba błędów	Liczba obiektów	Procent błędnych decyzji
5	40	12,5% (zbiór treningowy)
56	352	15,91% (zbiór testowy)

Wartość $MaxP$

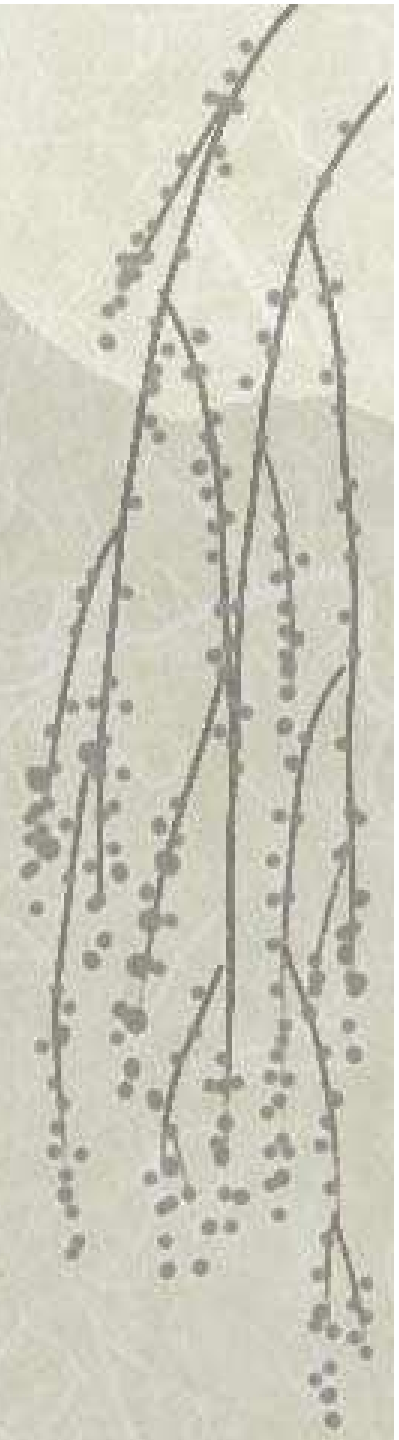
- ❖ Zbyt małe $MaxP$ – duży błąd z powodu zbyt dużego uogólnienia
- ❖ Zbyt duże $MaxP$ – duży błąd z powodu przetrenowania
- ❖ Nie ma jednej uniwersalnej wartości $MaxP$
- ❖ ALE: Dla określonego zbioru danych można automatycznie wyznaczyć najlepsze $MaxP$
 - *Metoda kros-walidacji*

Drzewa dla danych rzeczywistych

- ❖ Zbiory danych zawierać mogą atrybuty opisane wartościami ciągłymi
 - Przyspieszenie, rok produkcji, zużycie paliwa
- ❖ Rozgałęzienie na każde możliwe wartości?
 - Przetrenowanie!
 - Duża wartość p doprowadzi do usunięcia całych poziomów drzewa!
- ❖ Rozgałęzienie na przedziały wartości!

Przedziały wartości

- ❖ Dyskretyzacja wartości ciągłych
- ❖ Zamiast wartości posługujemy się nazwą/etykieta/symbolem przedziału
- ❖ Jak wyznaczać przedziały?



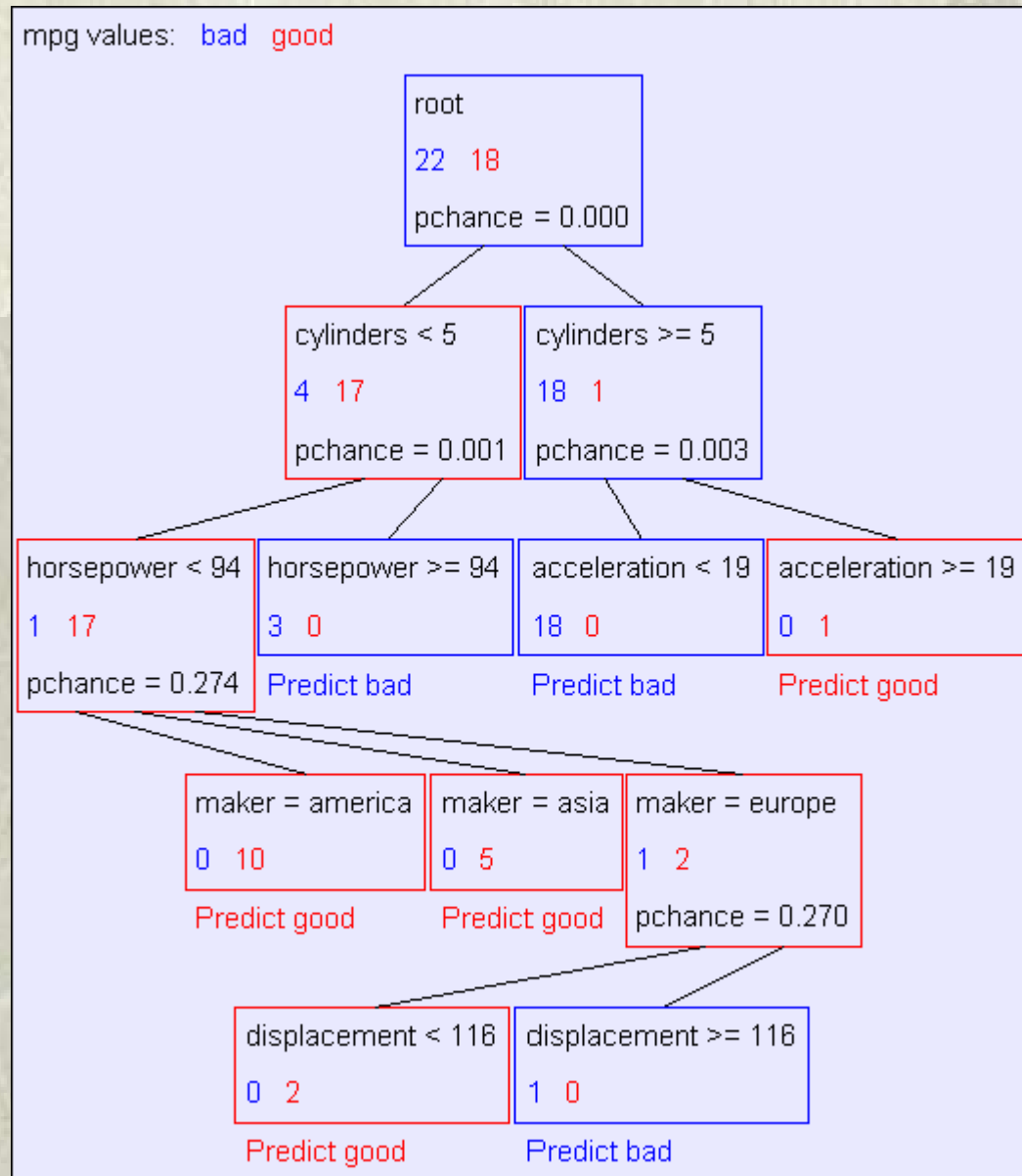
Przedziały wartości – zysk informacyjny

- ❖ Było: $IG(Y|X) = H(Y) - H(Y|X)$
- ❖ Niech: $IG(Y|X:t) = H(Y) - H(Y|X:t)$
 $H(Y|X:t) = H(Y|X < t) * P(Y|X < t) +$
 $+ H(Y|X \geq t) * P(Y|X \geq t)$
- ❖ $IG(Y|X:t)$ zysk informacyjny dla wartości Y pod warunkiem, że wiadomo, czy X jest większe czy mniejsze od t
- ❖ t - wartość dzieląca dziedzinę X na przedziały

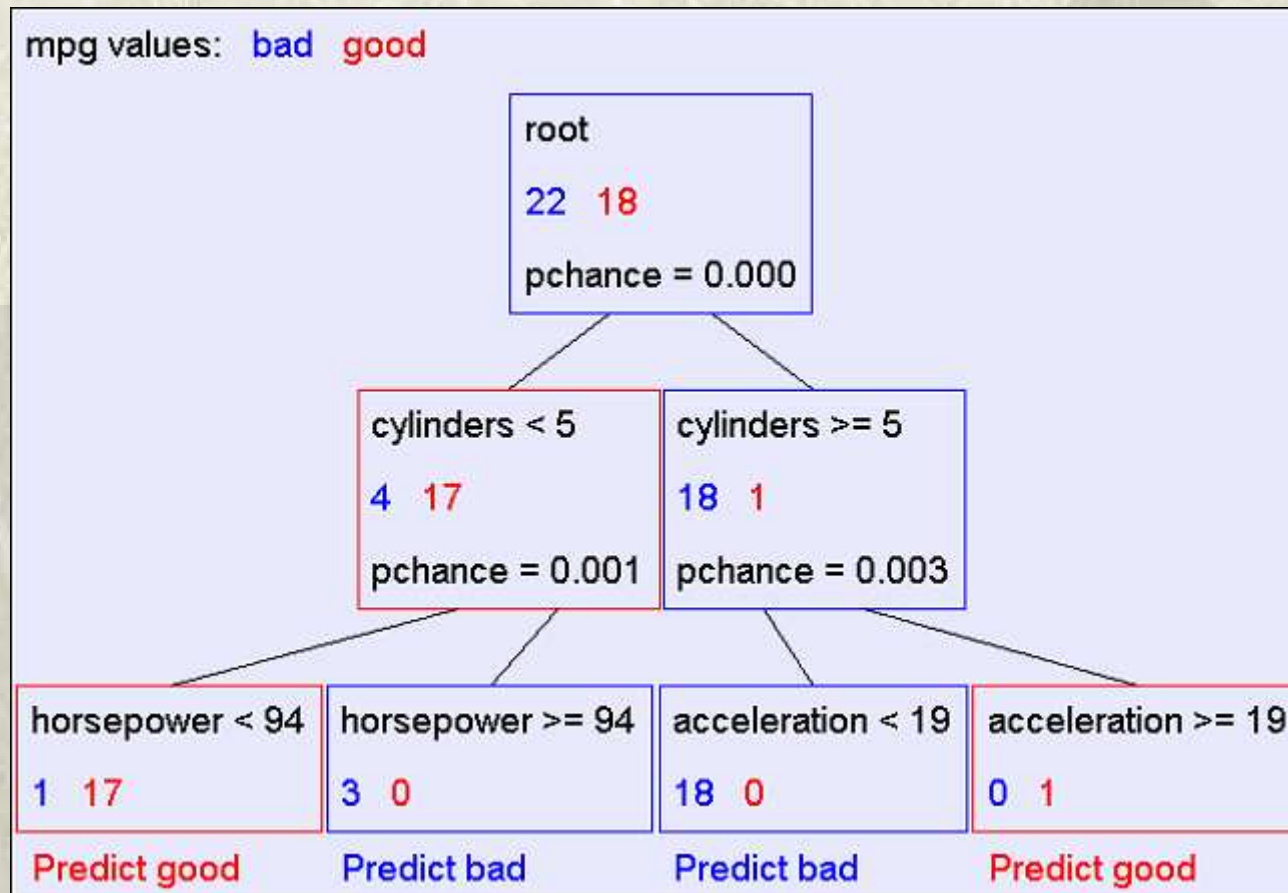
Przedziały wartości – zysk informacyjny

- ❖ Niech: $IG^*(Y|X) = \max_t(IG(Y|X:t))$
- ❖ t – miejsce podziału generujące największy IG
- ❖ W trakcie budowania drzewa atrybut X nadaje się na rozgałęzienie w zależności od jego wartości $IG^*(Y|X)$

Przykład – drzewo decyzyjne



Przykład – drzewo decyzyjne



Liczba błędów	Liczba obiektów	Procent błędnych decyzji
1	40	2,5% (zbiór treningowy)
53	352	15,06% (zbiór testowy)

Podsumowanie

- ❖ Drzewa decyzyjne:
 - Łatwe do interpretacji
 - Łatwe do implementacji
 - Łatwe do wykorzystania
 - Proste obliczeniowo
- ❖ Uwaga na przetrenowanie
- ❖ Realizują skutecznie zadanie klasyfikacji