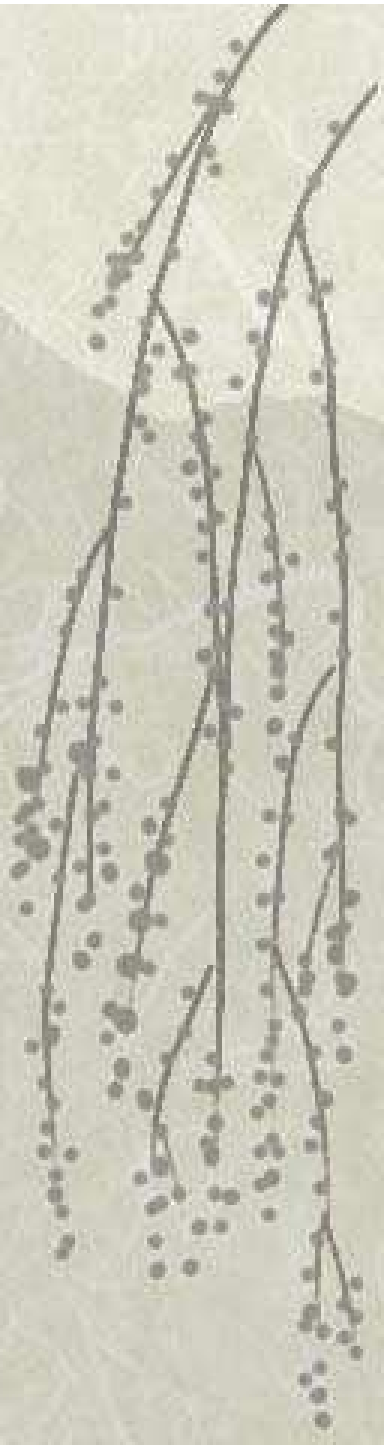
Drzewa Decyzyjne, cz.1

Inteligentne Systemy Decyzyjne

Katedra Systemów Multimedialnych

WETI, PG

Opracowanie: dr inż. Piotr Szczuko



Zadanie klasyfikacji

- Najważniejsza operacja w drążeniu danych (ang. *Data Mining*)
- Próba przewidywania wyniku (określania **kategorii**, klasyfikowania) na podstawie posiadanych parametrów opisujących obiekt
- **Kategoria**: atrybut symboliczny, przyjmuje dyskretne wartości, np. opis słowny.
- Przeciwieństwo to atrybut liczbowy, przyjmuje wartości ciągłe, to np. dowolne wielkości fizyczne

Zbiór danych

- Przykładowe dane ankiet oceny „stopy życiowej” (48,000 rekordów, 16 atrybutów opublikowane [Kohavi 1995])

Age	employmen	education	edun	marital	…	job	relation	race	gender	hours	country	wealth
…												
39	State_gov	Bachelors	13	Never_mar	…	Adm_cleric	Not_in_fam	White	Male	40	United_Sta	poor
51	Self_emp_	Bachelors	13	Married	…	Exec_man	Husband	White	Male	13	United_Sta	poor
39	Private	HS_grad	9	Divorced	…	Handlers_c	Not_in_fam	White	Male	40	United_Sta	poor
54	Private	11th	7	Married	…	Handlers_c	Husband	Black	Male	40	United_Sta	poor
...												
52	Self_emp_	HS_grad	9	Married	…	Exec_man	Husband	White	Male	45	United_Sta	rich
31	Private	Masters	14	Never_mar	…	Prof_speci	Not_in_fam	White	Female	50	United_Sta	rich
32	...											

Analiza danych

- Histogram?

Value	Frequency
Female	16192
Male	32650

Gender

Value	Frequency
Divorced	6633
Married_AF_spouse	37
Married	22379
Married_spouse_absent	628
Never_married	16117
Separated	1530
Widowed	1518

Marital Status

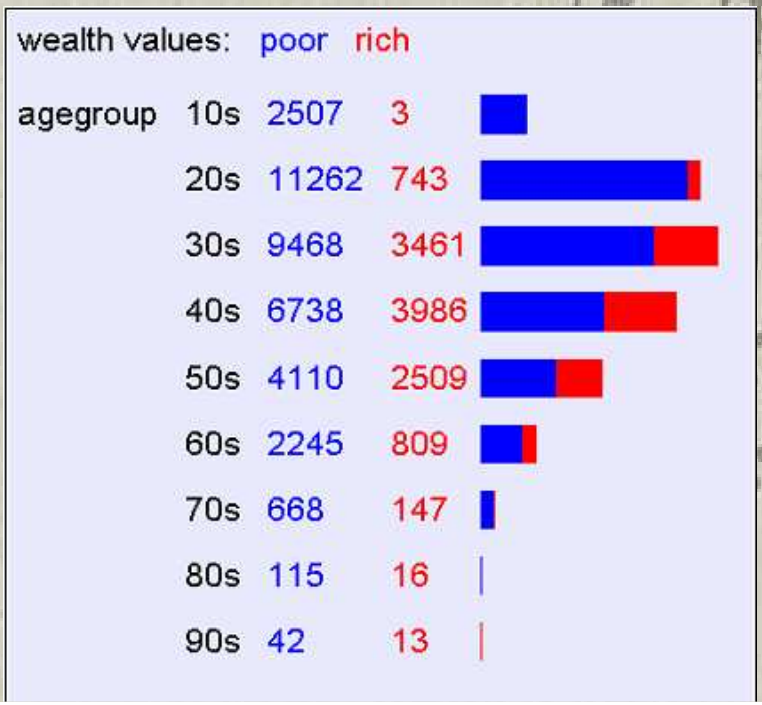
Tablice kontyngencji

- Kontyngencja „współzależność statystyczna między cechami, z których przynajmniej jedna jest cechą jakościową”
- Histogram to jednowymiarowa tablica kontyngencji
- k-wymiarowa tablica:
  - Wybierz k atrybutów ze zbioru danych: a_1, a_2, ..., a_k
  - Dla każdej możliwej kombinacji wartości...
    - $a_1=x_1$, $a_2=x_2$, …, $a_k=x_k$
  - ...określ jak często pojawia się ona w zbiorze danych

Przykład tablicy kontyngencji

- Dla każdej możliwej pary wartości atrybutów (wiek, zamożność) określamy liczbę wystąpień w zbiorze danych:

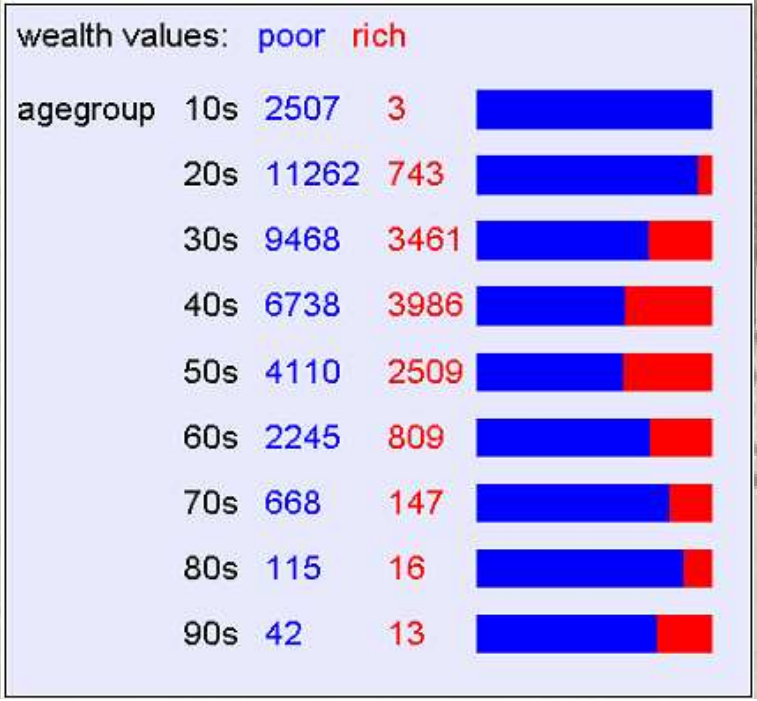
wealth values:		poor	rich
agegroup	10s	2507	3
	20s	11262	743
	30s	9468	3461
	40s	6738	3986
	50s	4110	2509
	60s	2245	809
	70s	668	147
	80s	115	16
	90s	42	13



Przykład tablicy kontyngencji

- Dla każdej możliwej pary wartości atrybutów (wiek, zamożność) określamy liczbę wystąpień w zbiorze danych:

wealth values:	poor	rich
agegroup 10s	2507	3
20s	11262	743
30s	9468	3461
40s	6738	3986
50s	4110	2509
60s	2245	809
70s	668	147
80s	115	16
90s	42	13

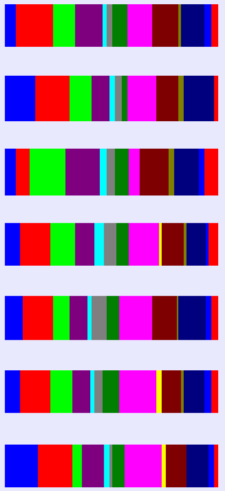


Przykład tablicy kontyngencji

- ... Tablica 2-wymiarowa, c.d.

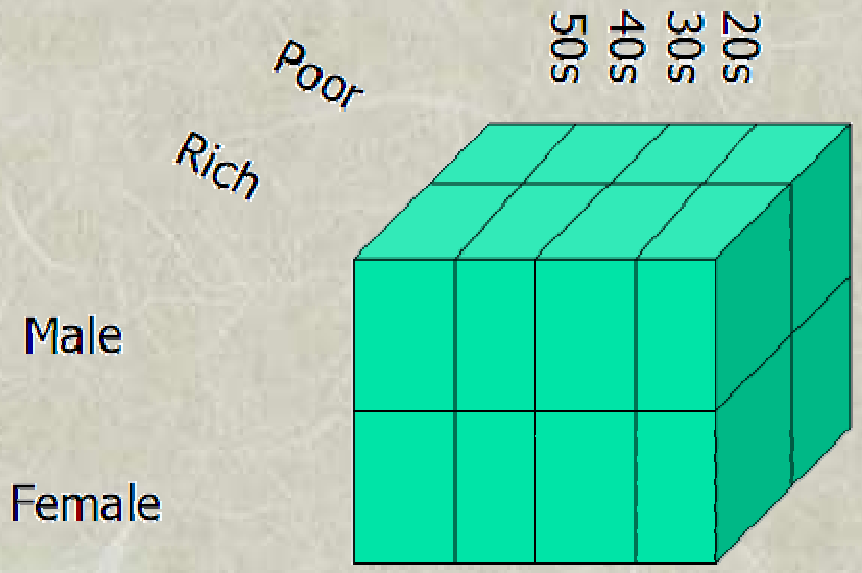
job values:	Adm_clerical	Craft_repair	Farming_fishing	Machine_op_inspct	Priv_house_serv	Protective_serv	Tech_support
MissingValue	Armed_Forces	Exec_managerial	Handlers_cleaners	Other_service	Prof_specialty	Sales	Transport_moving

marital																
	Divorced	270	1192	0	679	890	90	197	434	762	46	795	121	664	239	254
	Married_AF_spouse	5	6	0	4	3	1	1	1	5	0	4	1	5	0	1
	Married	928	1495	7	3818	3600	869	724	1469	1088	27	3182	583	2491	609	1489
	Married_spouse_absent	45	84	0	77	52	35	32	37	92	9	64	7	55	9	30
	Never_married	1242	2360	8	1301	1260	434	1029	872	2442	99	1849	237	1992	506	486
	Separated	97	224	0	160	126	23	63	123	275	21	145	23	146	48	56
	Widowed	222	250	0	73	155	38	26	86	259	40	133	11	151	35	39



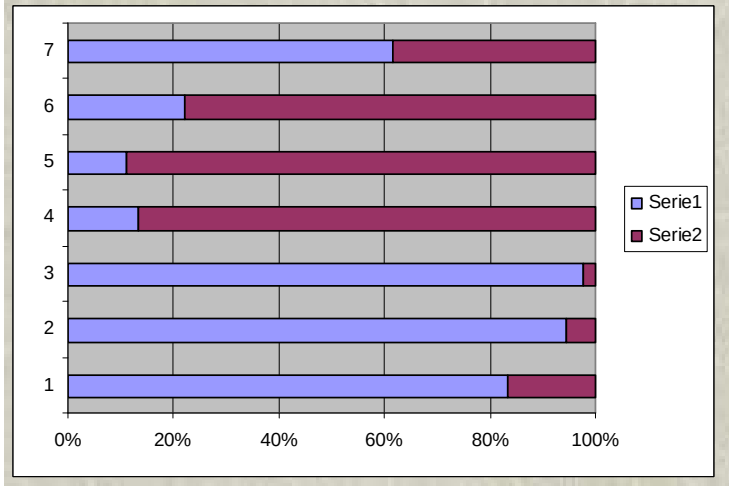
Przykład tablicy kontyngencji

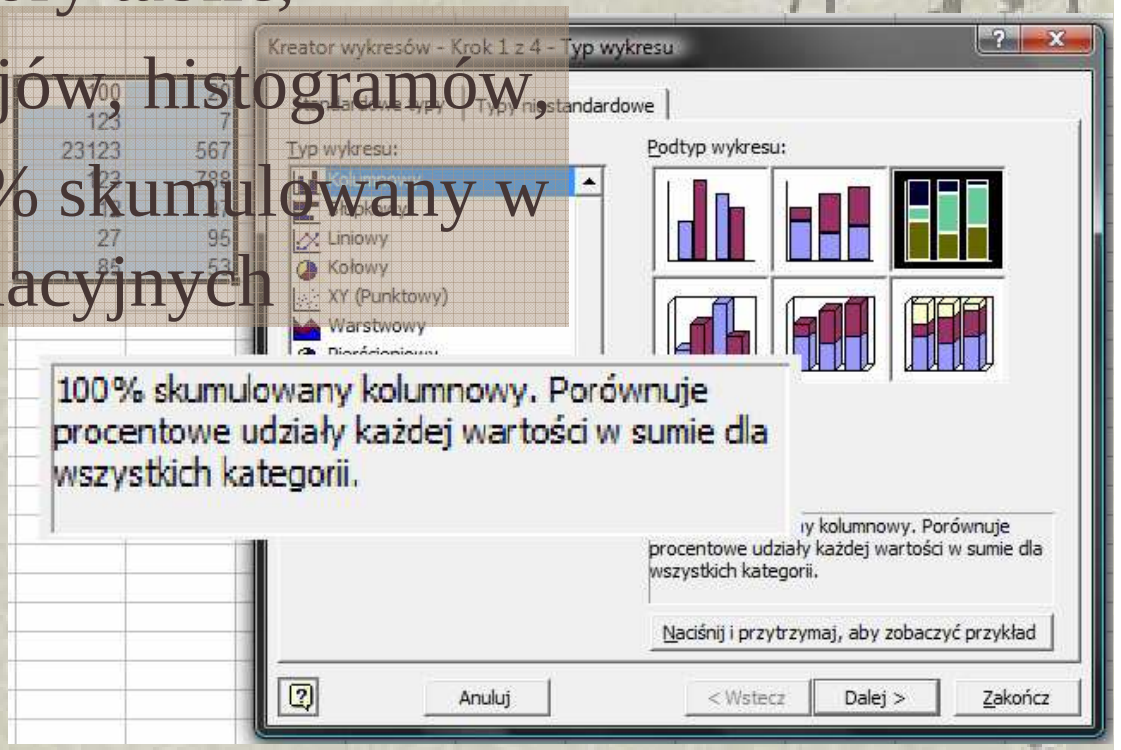
- ...Tablica 3-wymiarowa



Analiza danych, c.d.

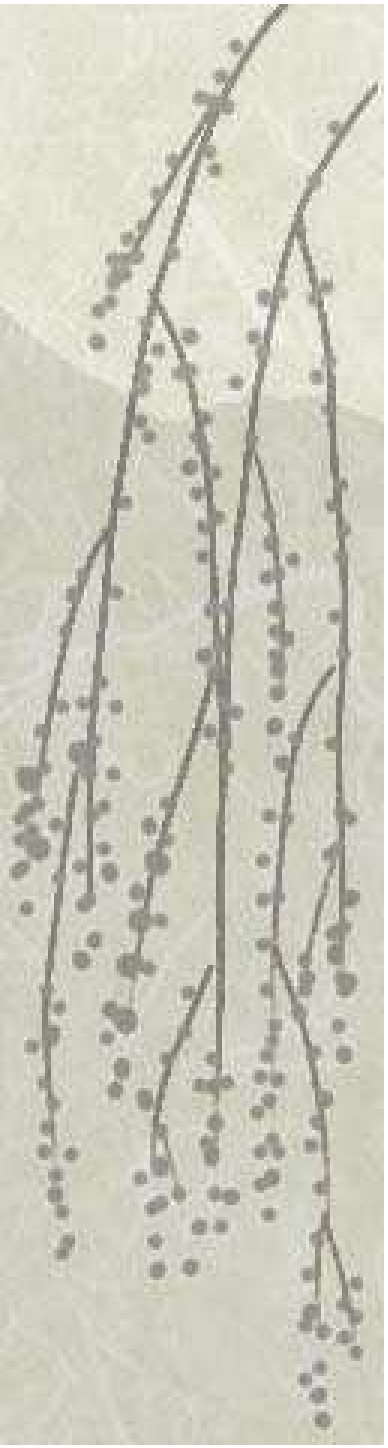
- Oprogramowanie typu OLAP (On-line Analytical Processing):
  - „Klikalne” kreatory tablic,
  - Podgląd przekrojów, histogramów,
  - Np. wykres 100% skumulowany w arkuszach kalkulacyjnych





Tablice kontyngencji

- Wady?
- Dla 16 atrybutów w zbiorze danych:
  - Ile jest tablic 1-wymiarowych?
  - Ile jest tablic 2-wymiarowych?
  - Ile jest tablic 3-wymiarowych?
  - Ile jest tablic 3-wymiarowych dla 100 atrybutów?



Tablice kontyngencji

- Wady?
- Dla 16 atrybutów w zbiorze danych:
  - Ile jest tablic 1-wymiarowych? **16**
  - Ile jest tablic 2-wymiarowych? 16*15/2=**120**
  - Ile jest tablic 3-wymiarowych? **560**
  - Ile jest tablic 3-wymiarowych dla 100 atrybutów? **161700**

Drążenie danych

- Automatyczne poszukiwanie związków i zależności zawartych w zbiorach danych.

- Jakie zależności są interesujące?
- Jakie są znaczące tylko pozornie?
- Jak je wykorzystywać?

Istotność informacji

- Z teorii informacji → „jakość informacji”
  - (nie-)nadmiarowa?
  - (nie-)uporządkowana?
  - (nie-)istotna?

  - Wartość ENTROPII - najmniejsza średnia **ilość informacji** potrzebna do zakodowania faktu zajścia pewnego zdarzenia (ze zbioru zdarzeń o danych prawdopodobieństwach)

Entropia

- Przykład prawdopodobieństw zdarzeń:
  - P(X=A) = 1/4
  - P(X=B) = 1/4
  - P(X=C) = 1/4
  - P(X=D) = 1/4
- Ciąg zdarzeń ABBDCADC przesyłamy zakodowany binarnie, każde na 2 bitach, np.
  - A = 00, B = 01, C = 10, D = 11
  - 0001011110001110

Entropia

- Niech prawdopodobieństwa będą **różne**:
  - P(X=A) = 1/2
  - P(X=B) = 1/4
  - P(X=C) = 1/8
  - P(X=D) = 1/8
- Możliwe jest takie zakodowanie binarne zdarzeń, że **średnio** potrzebne będzie tylko 1.75 bita na zdarzenie!
  - A = 0, B = 10, C = 110, D = 111

Entropia

- Przypadek ogólny:
  - Zmienna losowa X przyjmuje wartości V_1, V_2, ..., V_m z P równymi p_1, p_2, ..., p_m.
  - Najmniejsza możliwa średnia liczba bitów na symbol potrzebnych do przetransmitowania ciągu symboli o dystrybucji losowej X:
  - $H(X) = -p_1 log_2 p_1 - p_2 log_2 p_2 - ... - p_m log_2 p_m = -\Sigma_{j=1...m} p_j log_2 p_j$
  - H(X) to **ENTROPIA** X

Entropia

- Policzmy dla:
  - P(X=A) = 1/2
  - P(X=B) = 1/4
  - P(X=C) = 1/8
  - P(X=D) = 1/8
  - $H(X) = -p_1 log_2 p_1 - p_2 log_2 p_2 - \ldots - p_m log_2 p_m = ?$

Entropia

- Duża entropia oznacza równomierną dystrybucję, płaski histogram, rozrzucenie wartości w całym przedziale, **trudniejsze do przewidzenia**
- Mała entropia oznacza zmienną dystrybucję, pofalowany histogram, skupiska i luki w przedziale wartości, **łatwiejsze do przewidzenia**

Entropia warunkowa

- Próba przewidywania wartości wyjściowej Y pod warunkiem znania wartości wejściowej X:
- Prawdopodobieństwa i p. warunkowe:
  - P(Y = Yes) = 0.5
  - P(X = Math & Y = No) = 0.25
  - P(X = Math) = 0.5
  - P(Y = Yes | X = History) = 0

Zauważmy:

  - H(X) = 1.5
  - H(Y) = 1 *(dlaczego?)*

- Entropia warunkowa H(Y|X=v) – entropia liczona po tych tylko Y dla których X=v.
  - Ile wynoszą H(Y|X=Math), H(Y|X=CS), H(Y|X=History) ?

X	Y
Math	Yes
History	No
CS	Yes
Math	No
Math	No
CS	Yes
History	No
Math	Yes

Entropia warunkowa

- H(Y |X) = średnia entropia warunkowa Y-ka
- = jeżeli wybiorę wiersz ze zbioru danych losowo, to jaka będzie entropia warunkowa Y-ka, jeżeli poznam w tym wierszu wartość X
- = oczekiwana liczba bitów do przesłania Y, pod warunkiem, że nadawca i odbiorca znają X'y
- $= \Sigma_j P(X=v_j) H(Y \mid X = v_j)$

Entropia warunkowa

v_j	$Prob(X=v_j)$	$H(Y \mid X = v_j)$
Math	0.5	1
History	0.25	0
CS	0.25	0

X	Y
Math	Yes
History	No
CS	Yes
Math	No
Math	No
CS	Yes
History	No
Math	Yes

- $H(Y \mid X) = \Sigma_j P(X=v_j) H(Y \mid X = v_j)$
- $H(Y \mid X) = 0.5*1 + 0.25*0 + 0.25*0 = 0.5$

Zysk informacyjny

- IG(ang. *Information Gain*)
- IG(Y|X) = Jeżeli muszę przesłać Y, ile oszczędzę bitów średnio, jeżeli obie strony znać będą X?
- IG(Y|X) = H(Y) - H(Y | X)
- Przykład:
  - H(Y) = 1
  - H(Y|X) = 0.5
  - IG(Y|X) = 1 – 0.5 = 0.5

Przykłady

- IG(Y|X) = H(Y) - H(Y | X)

wealth values: poor rich

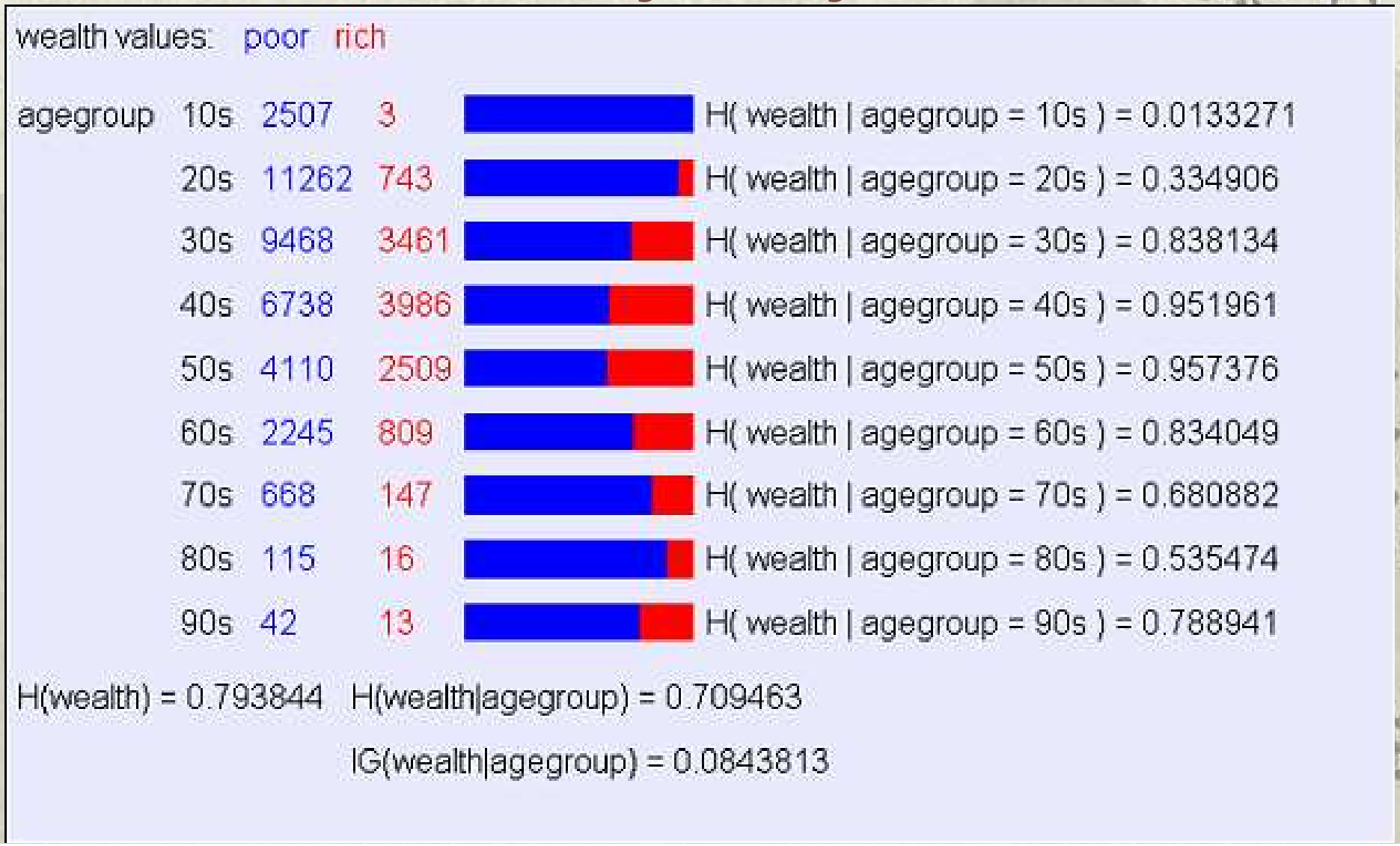
gender Female 14423 1769  H(wealth | gender = Female) = 0.497654

Male 22732 9918  H(wealth | gender = Male) = 0.885847

H(wealth) = 0.793844 H(wealth|gender) = 0.757154

IG(wealth|gender) = 0.0366896

Przykłady


wealth values: poor rich
agegroup 10s 2507 3 H(wealth | agegroup = 10s) = 0.0133271
20s 11262 743 H(wealth | agegroup = 20s) = 0.334906
30s 9468 3461 H(wealth | agegroup = 30s) = 0.838134
40s 6738 3986 H(wealth | agegroup = 40s) = 0.951961
50s 4110 2509 H(wealth | agegroup = 50s) = 0.957376
60s 2245 809 H(wealth | agegroup = 60s) = 0.834049
70s 668 147 H(wealth | agegroup = 70s) = 0.680882
80s 115 16 H(wealth | agegroup = 80s) = 0.535474
90s 42 13 H(wealth | agegroup = 90s) = 0.788941
H(wealth) = 0.793844 H(wealth|agegroup) = 0.709463
IG(wealth|agegroup) = 0.0843813

Przydatność Zysku Informacyjnego

- Przypuśćmy, że chcemy określić, czy ktoś zaliczy ISD:
- IG(Zaliczenie | KolorWłosów) = 0.01
- IG(Zaliczenie | Obecność na wykładach) = 0.55
- IG(Zaliczenie | Czas nauki) = 0.25
- IG(Zaliczenie | OstatniaCyfraNrIndeksu) = 0.00001
- IG mówi jak interesująca będzie wybrana 2-wymiarowa tablica kontyngencji

Drzewo Decyzyjne

- drzewiasta struktura opisująca zbiór atrybutów, których wartości należy testować w zadanej kolejności w celu przewidzenia wartości wyjściowej.
- Aby określić, które z nich testować na początku, należy znaleźć atrybut nieprzetestowany o najwyższej wartości IG...
- ...następnie proces powtórzyć, budując tak całe drzewo

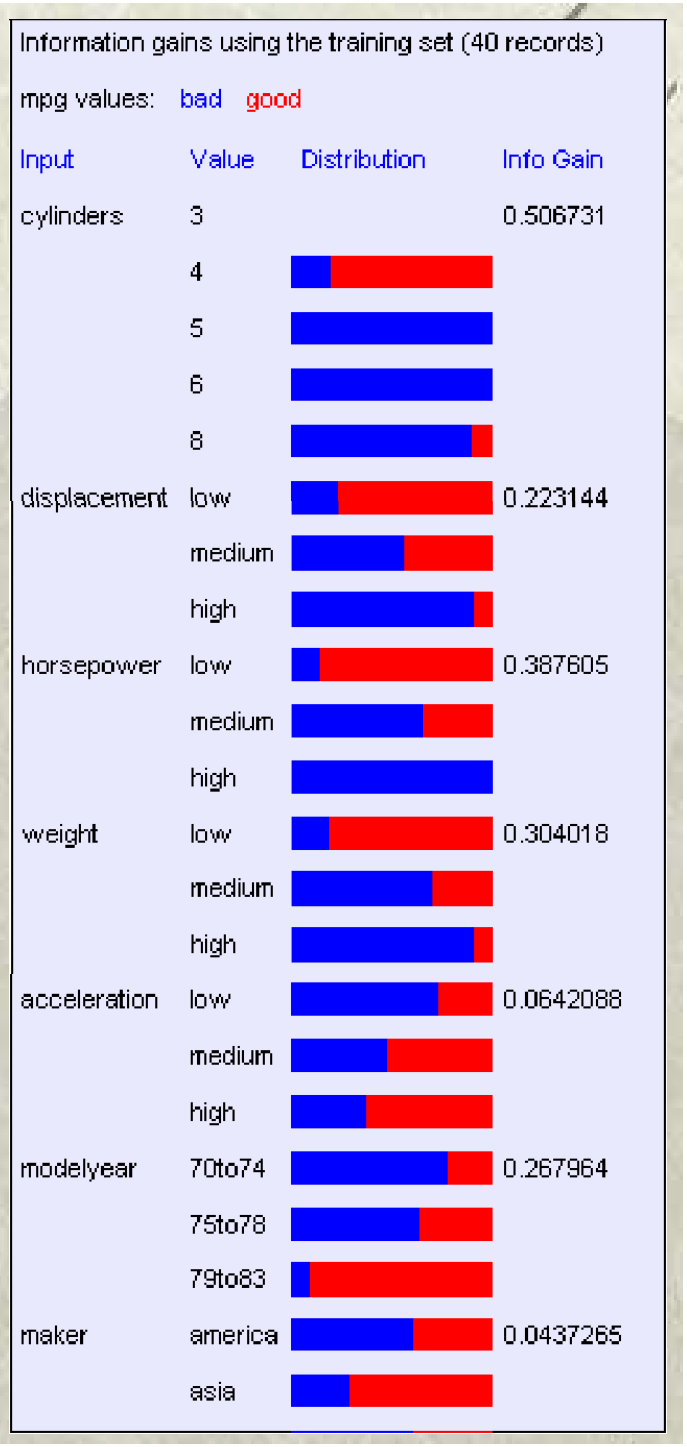
Przykład motoryzacyjny :)

Zbiór danych

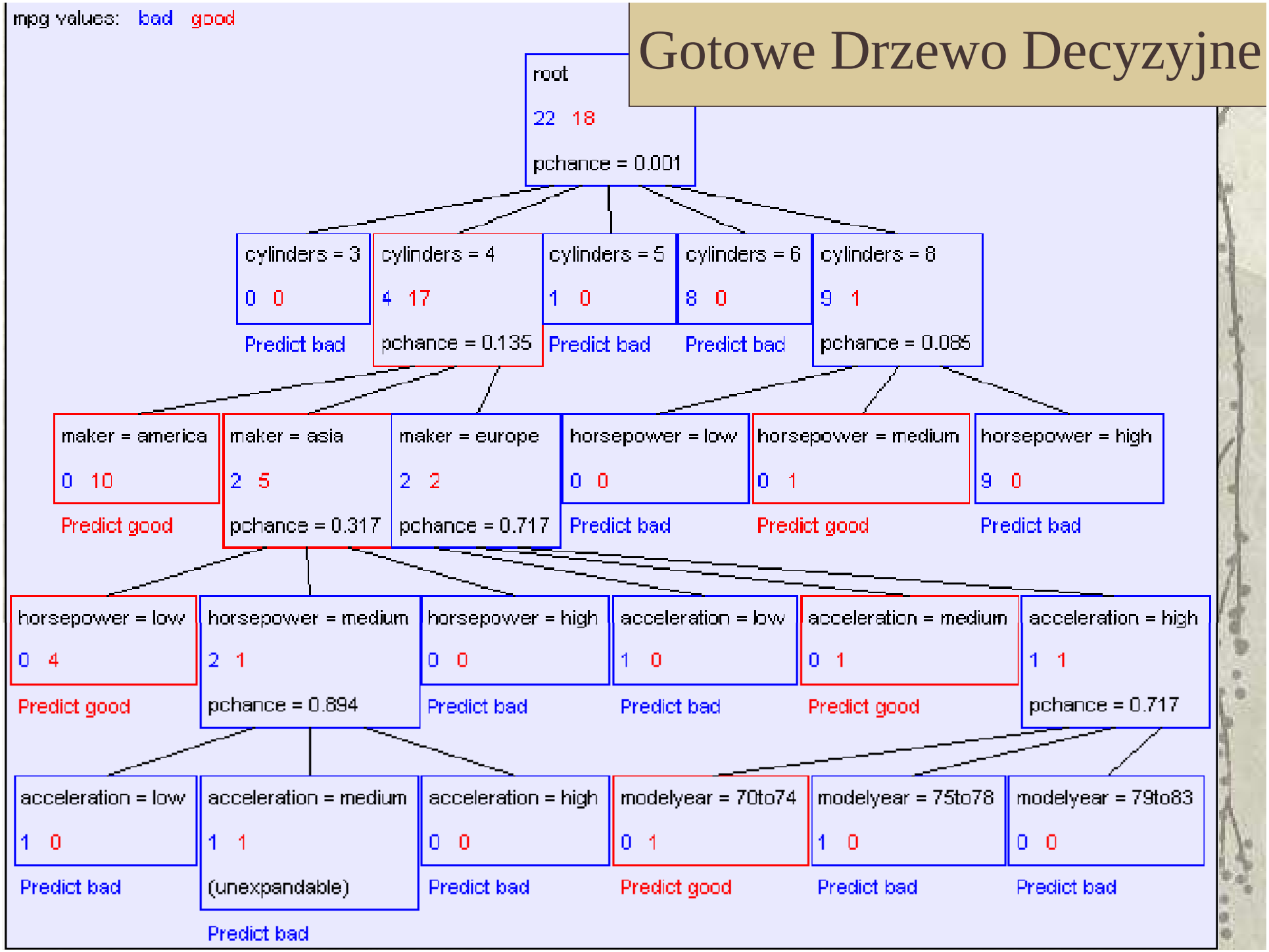
mpg	cylinders	displacement	horsepower	weight	acceleration	modelyear	maker
good	4	low	low	low	high	75to78	asia
bad	6	medium	medium	medium	medium	70to74	america
bad	4	medium	medium	medium	low	75to78	europe
bad	8	high	high	high	low	70to74	america
bad	6	medium	medium	medium	medium	70to74	america
bad	4	low	medium	low	medium	70to74	asia
bad	4	low	medium	low	low	70to74	asia
bad	8	high	high	high	low	75to78	america
:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:
bad	8	high	high	high	low	70to74	america
good	8	high	medium	high	high	79to83	america
bad	8	high	high	high	low	75to78	america
good	4	low	low	low	low	79to83	america
bad	6	medium	medium	medium	high	75to78	america
good	4	medium	low	low	low	79to83	america
good	4	low	low	medium	high	79to83	america
bad	8	high	high	high	low	70to74	america
good	4	low	medium	low	medium	75to78	europe
bad	5	medium	medium	medium	medium	75to78	europe

Wartości IG

- Aby określić, które z nich testować na początku, należy znaleźć atrybut nieprzetestowany o najwyższej wartości IG...
- ...następnie proces powtórzyć, budując tak całe drzewo



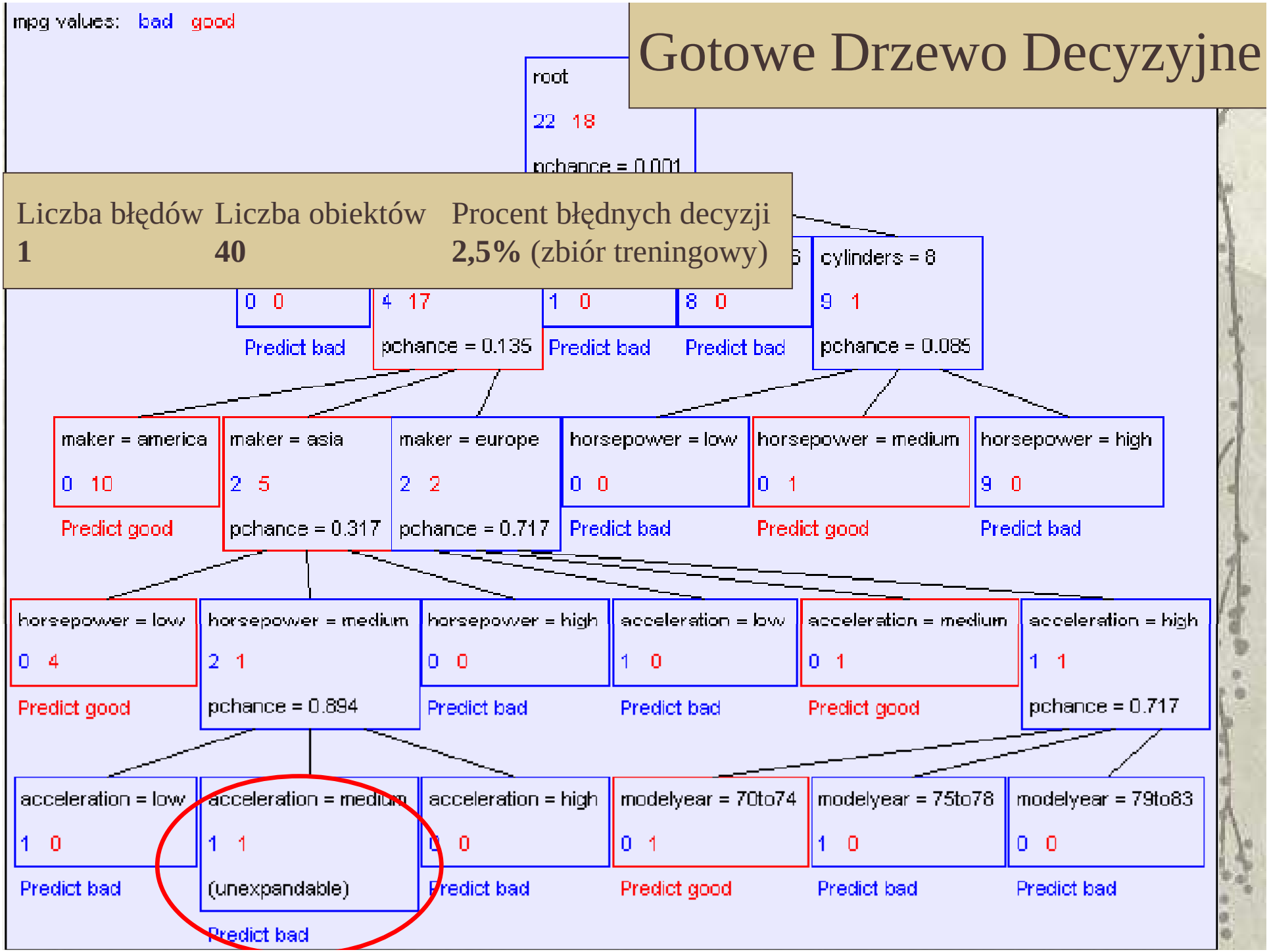
Gotowe Drzewo Decyzyjne



Błąd zbioru danych (treningowy)

- Dla każdego rekordu w tablicy przejdź kolejne etapy drzewa i sprawdź, czy wskazana „w liściu” wartość zgadza się z daną wartością wyjściową w tabeli
- Jaka liczba tych wartości się nie zgadza?
  - Jest ona nazywana błędem treningowym (im jest mniejszy tym lepiej)

Gotowe Drzewo Decyzyjne



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

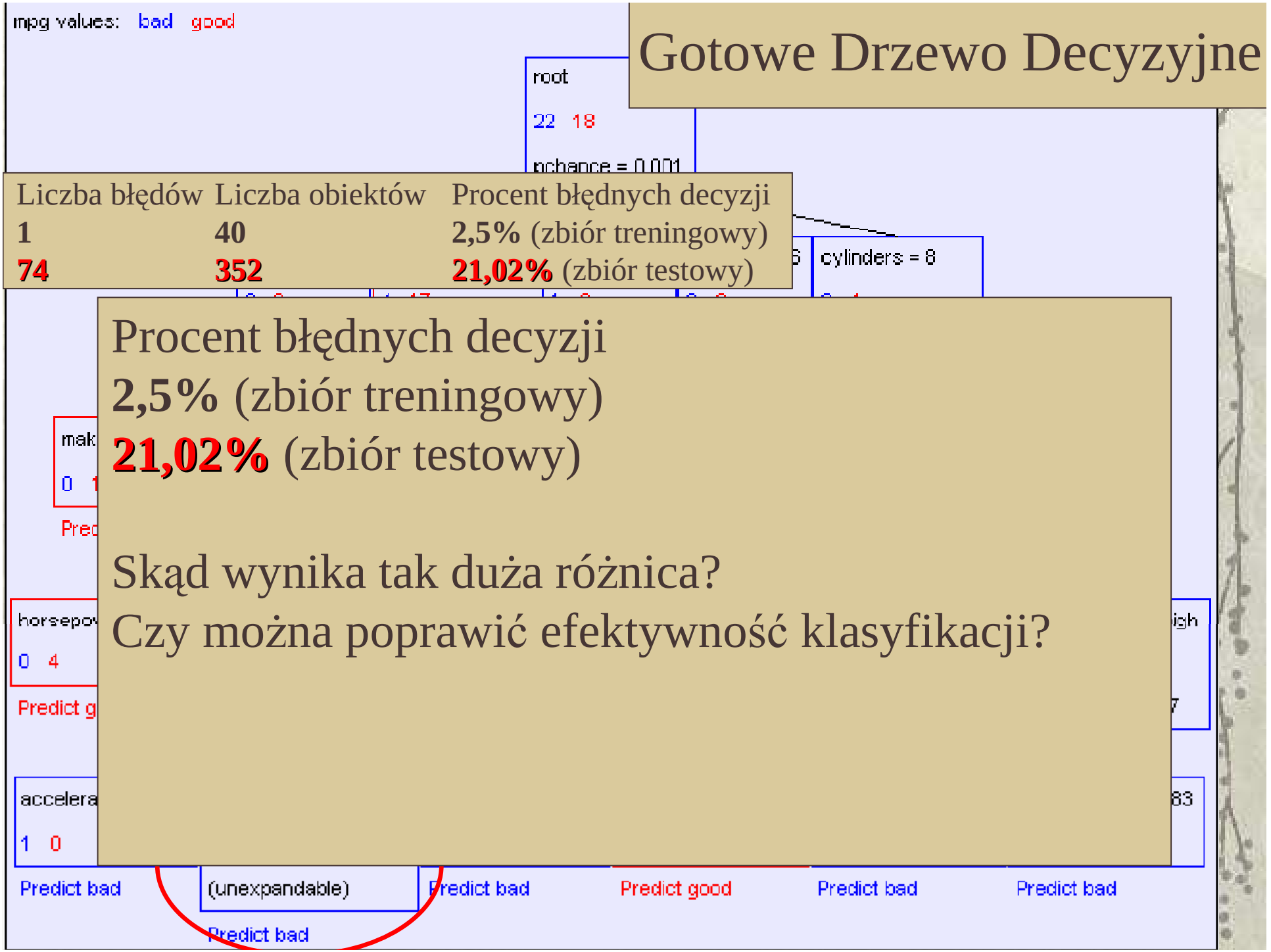
Klasyfikacja nieznanych obiektów

- Naszym celem nie jest klasyfikacja obiektów już występujących w tabeli (i testowanie skuteczności tej klasyfikacji)
- Drzewa decyzyjne stosowane są do klasyfikacji obiektów nieznanych.
- Stworzenie aparatu decyzyjnego, który w przyszłości podejmie właściwe decyzje.

Zbiór testowy

- Zbiór danych – dane treningowe = dane testowe
- Wydzielenie przed tworzeniem drzewa danych, które posłużą do przetestowania trafności klasyfikacji
- Sposób zasymulowania tego, co może się wydarzyć w przyszłości

Gotowe Drzewo Decyzyjne



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

Procent błędnych decyzji

2,5% (zbiór treningowy)

21,02% (zbiór testowy)

Skąd wynika tak duża różnica?

Czy można poprawić efektywność klasyfikacji?