GDAŃSK UNIVERSITY OF TECHNOLOGY

Eksploracja danych

KLASYFIKACJA I REGRESJA cz. 1

Wojciech Waloszek

wowal@eti.pg.gda.pl

Teresa Zawadzka

tegra@eti.pg.gda.pl

Katedra Inżynierii Oprogramowania
Wydział Elektroniki, Telekomunikacji i Informatyki
Politechnika Gdańska

GDAŃSK UNIVERSITY OF TECHNOLOGY

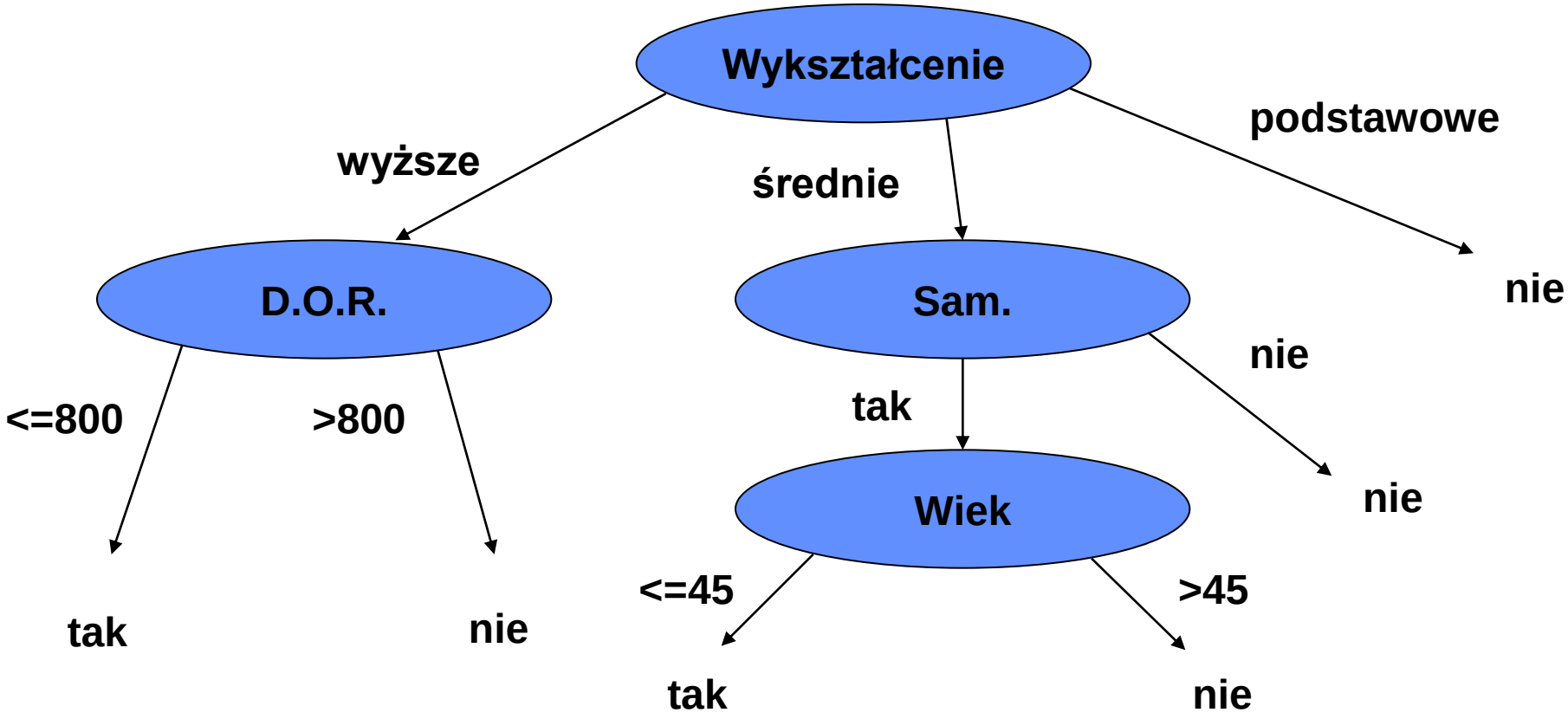
Budowa drzew decyzyjnych

- Drzewa decyzyjne to najpopularniejsza forma klasyfikatorów,

- Najczęściej budowane są metodą zstępującą, na zasadzie podejścia naturalnego dla drzew podejścia divide-and-conquer

GDAŃSK UNIVERSITY OF TECHNOLOGY

Przykład drzewa decyzyjnego



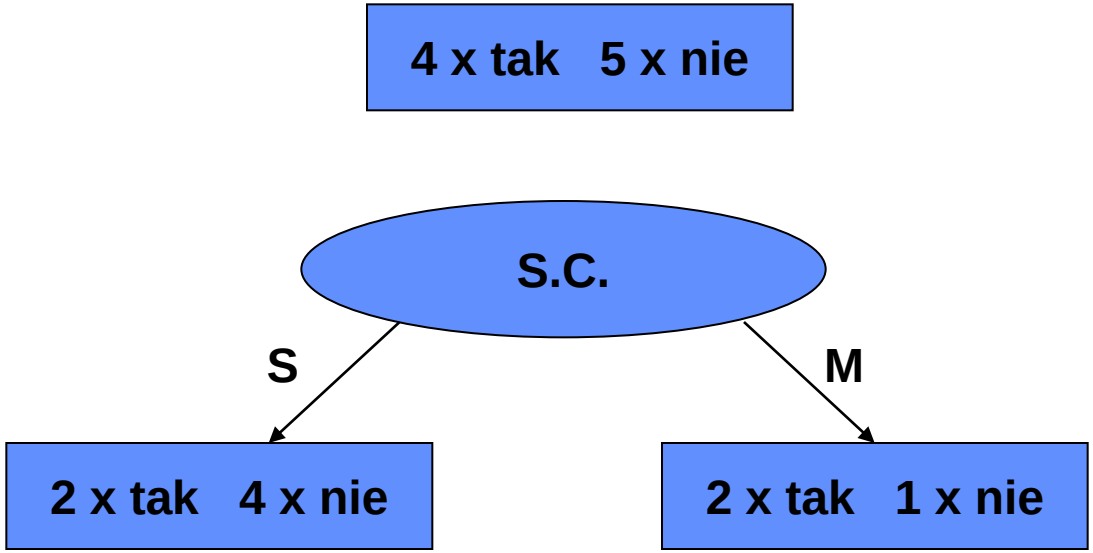
GDAŃSK UNIVERSITY OF TECHNOLOGY

Budowa drzewa

S.C.	D.O.R.	Wiek	Wykształcenie	Sam.	Z.K.
S	800	32	wyższe	tak	tak
S	1200	35	średnie	tak	tak
S	700	26	podstawowe	nie	nie
M	600	45	wyższe	nie	tak
M	650	38	średnie	tak	tak
S	900	28	wyższe	nie	nie
S	1100	65	średnie	tak	nie
M	500	22	średnie	nie	nie
S	800	43	podstawowe	tak	nie

1. **Mamy wyróżniony atrybut decyzyjny, wyznaczający *klasy***
2. **Na każdym poziomie drzewa wybieramy jeden z pozostałych atrybutów, najlepszy pod kątem dyskryminowania klas**
3. **Rozpoczynamy od pustego drzewa wyznaczając korzeń**

Dobór atrybutu



Czy podział pod względem wartości atrybutu S.C. jest korzystny?

I w jakiej mierze?

GDAŃSK UNIVERSITY OF TECHNOLOGY

Miara jakości podziału

- Jedną z miar jakości podziału jest *przyrost zawartości informacji*

- Przyrost zawartości informacji jest określony jako różnica zawartości *informacji* w dzielonym zbiorze przykładów a *entropią* zastosowanego podziału (*testu*).

GDAŃSK UNIVERSITY OF TECHNOLOGY

Miara jakości podziału – wzory

$$I(P) = \sum_{d \in C} -\frac{|P^d|}{|P|} \log_2 \frac{|P^d|}{|P|}$$

$$E_t(P) = \sum_{r \in R_t} \frac{|P_{tr}|}{|P|} I(P_{tr})$$

$I(P)$ – zawartość informacyjna zbioru przykładów P

C – zbiór klas wyznaczony przez atrybut decyzyjny

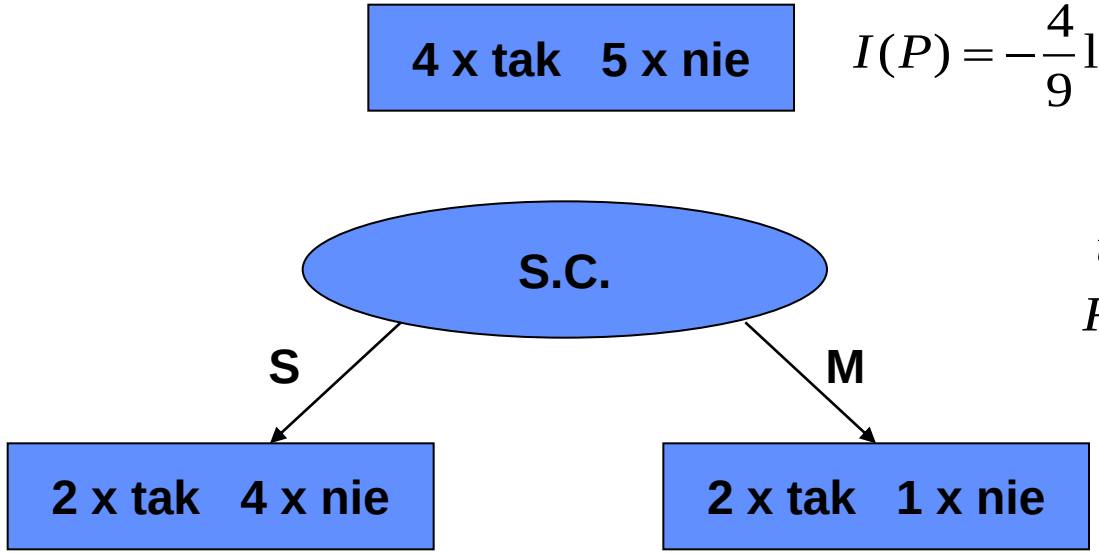
P^d – podzbiór tych przykładów ze zbioru P, które należą do klasy d

$E_t(P)$ – entropia testu t dla zbioru przykładów P

R_t – zbiór możliwych wyników testu t

P_{tr} – podzbiór tych przykładów ze zbioru P, które dają dla testu t wynik r

Dobór atrybutu - przykład



$$I(P) = -\frac{4}{9}\log_2\frac{4}{9} - \frac{5}{9}\log_2\frac{5}{9} \approx 0,99$$

$$t = "S.C. = ?"$$

$$R_t = \{S, M\}$$

$$I(P_{S.C.=S}) = -\frac{2}{6}\log_2\frac{2}{6} - \frac{4}{6}\log_2\frac{4}{6} \approx 0,92$$

$$I(P_{S.C.=M}) = -\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3} \approx 0,92$$

$$E_t(P) \approx \frac{6}{9}\cdot 0,92 + \frac{3}{9}\cdot 0,92 \approx 0,92$$

$$g_t(P) \approx 0,99 - 0,92 \approx 0,07$$

GDAŃSK UNIVERSITY OF TECHNOLOGY

Dobór atrybutu – przykład (2)

4 x tak 5 x nie

$$I(P) = -\frac{4}{9}\log_2\frac{4}{9} - \frac{5}{9}\log_2\frac{5}{9} \approx 0,99$$

$$t = "Wykształcenie = ?"$$

$$R_t = \{wyższe, średnie, podstawowe\}$$

Wykształcenie

wyższe

średnie

podstawowe

2 x tak 1 x nie

2 x tak 2 x nie

0 x tak 2 x nie

$$I(P_{Wykształcenie=wyższe}) \approx 0,92$$

$$I(P_{Wykształcenie=średnie}) = 1$$

$$I(P_{Wykształcenie=podstawowe}) = 0$$

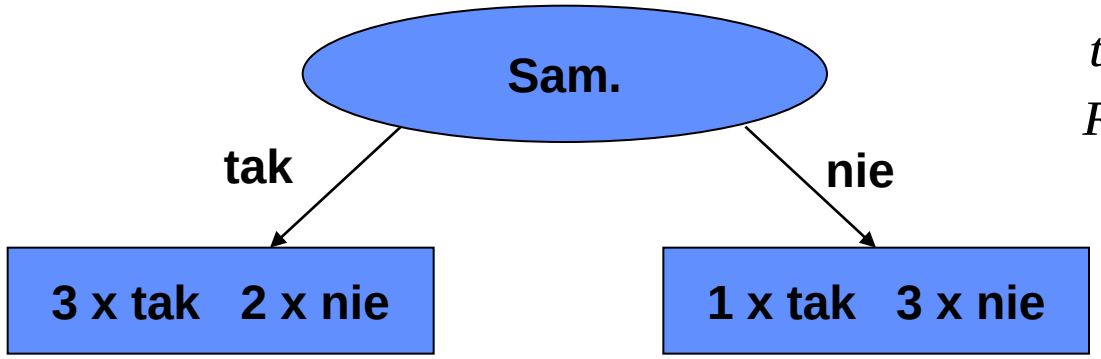
$$E_t(P) \approx \frac{3}{9}\cdot 0,92 + \frac{4}{9}\cdot 1 + \frac{2}{9}\cdot 0 \approx 0,75$$

$$g_t(P) \approx 0,99 - 0,75 \approx 0,24$$

Dobór atrybutu – przykład (3)

4 x tak 5 x nie

$$I(P) = -\frac{4}{9}\log_2\frac{4}{9} - \frac{5}{9}\log_2\frac{5}{9} \approx 0,99$$



$$t = "Sam. = ?"$$
$$R_t = \{tak, nie\}$$

$$I(P_{Sam.=tak}) \approx 0,97$$

$$I(P_{Sam.=nie}) \approx 0,81$$

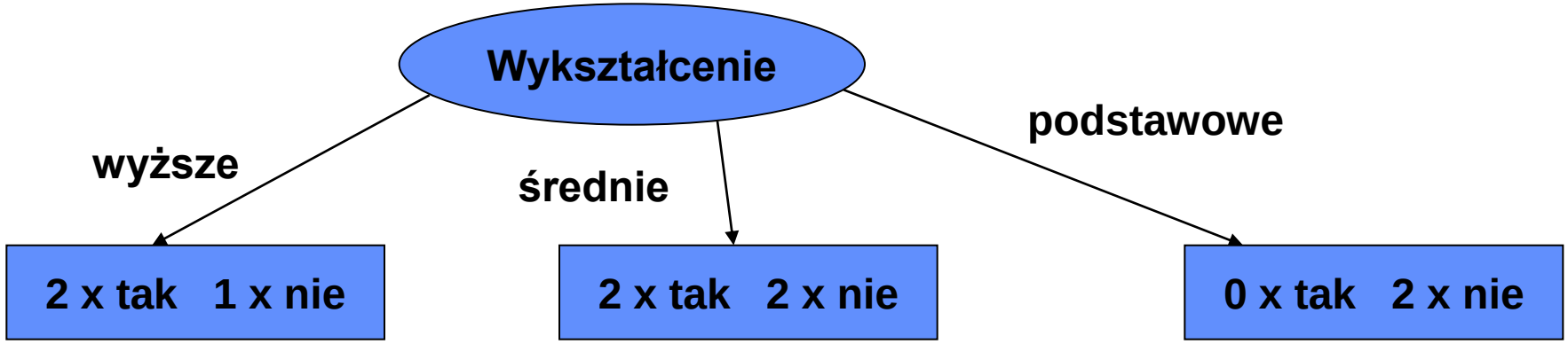
$$E_t(P) \approx \frac{5}{9}\cdot 0,97 + \frac{4}{9}\cdot 0,81 \approx 0,90$$

$$g_t(P) \approx 0,99 - 0,90 \approx 0,1$$

GDAŃSK UNIVERSITY OF TECHNOLOGY

Dobór atrybutu – przykład (4)

- Najwyższy zysk informacji (0,24) osiągnął atrybut *Wykształcenie* i on zostaje zapisany w korzeniu drzewa decyzyjnego



GDAŃSK UNIVERSITY OF TECHNOLOGY

Divide-and-conquer

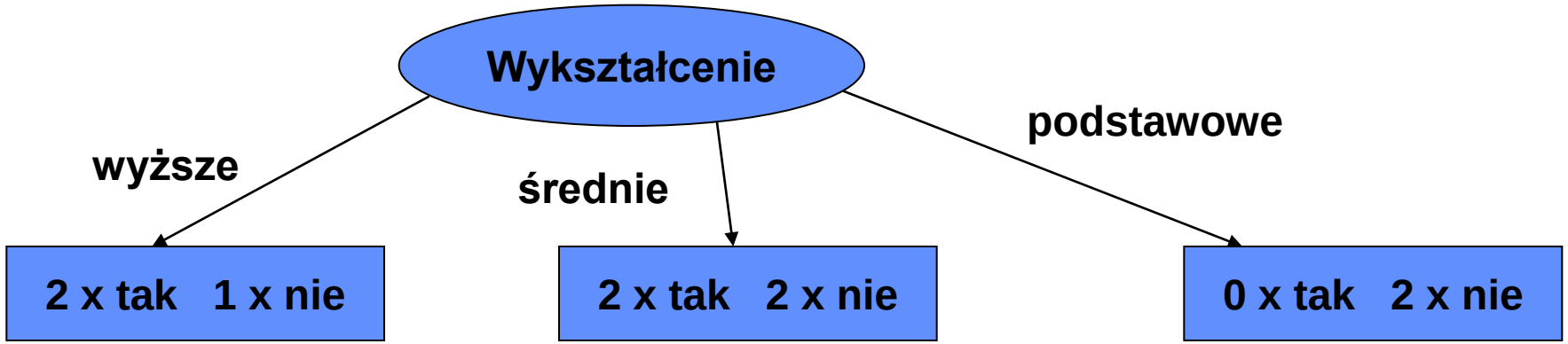
S.C.	D.O.R.	Wiek	Wykształcenie	Sam.	Z.K.
S	800	32	wyższe	tak	tak
S	1200	35	średnie	tak	tak
S	700	26	podstawowe	nie	nie
M	600	45	wyższe	nie	tak
M	650	38	średnie	tak	tak
S	900	28	wyższe	nie	nie
S	1100	65	średnie	tak	nie
M	500	22	średnie	nie	nie
S	800	43	podstawowe	tak	nie

1. **Zbiór przykładów został podzielony na trzy części**
2. **Dla każdej z części może zostać zastosowany ten sam algorytm dalszego działania**

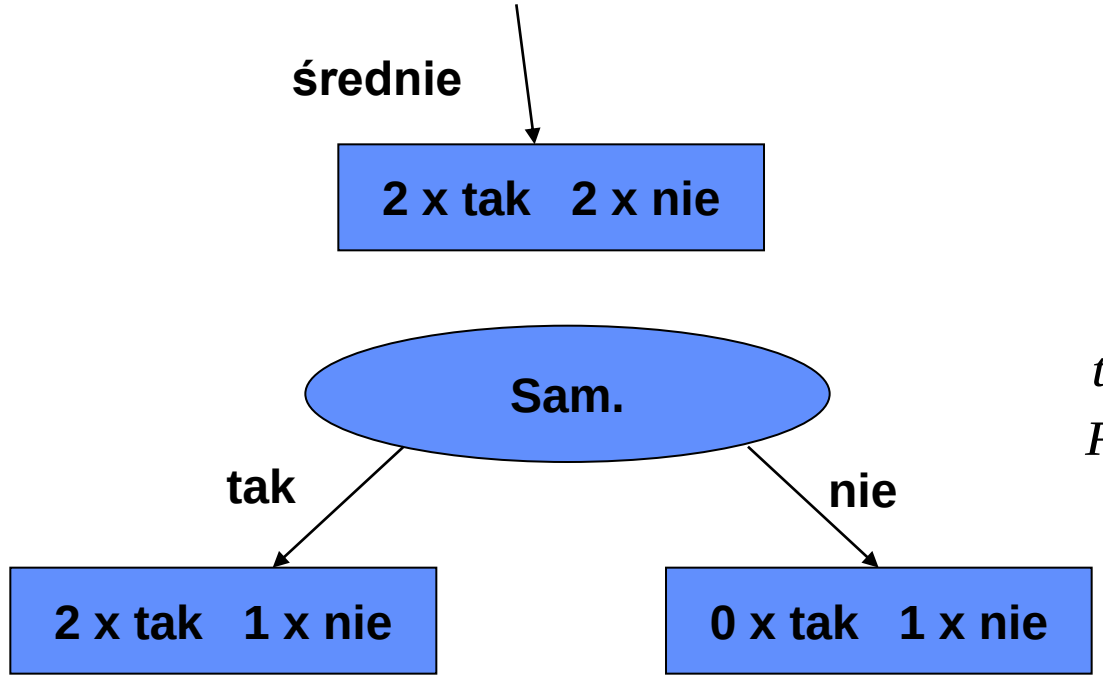
GDAŃSK UNIVERSITY OF TECHNOLOGY

Dalsza budowa drzewa

- Wzdłuż prawej gałęzi drzewa nie trzeba już rozbudowywać



Dalsza budowa drzewa (2)



$$I(P) = 1$$

$$t = "Sam. = ?"$$

$$R_t = \{tak, nie\}$$

$$I(P_{Sam.=tak}) \approx 0,92$$

$$I(P_{Sam.=nie}) \approx 0$$

$$E_t(P) \approx \frac{3}{4} \cdot 0,92 + \frac{1}{4} \cdot 0 \approx 0,69$$

$$g_t(P) \approx 1 - 0,69 \approx 0,31$$

GDAŃSK UNIVERSITY OF TECHNOLOGY

Atrybuty numeryczne

- Do tej pory zakładaliśmy użycie tylko atrybutów nominalnych,
- W trakcie budowy drzewa wykorzystywane mogą być też atrybuty numeryczne,
- Tutaj przedstawimy zasadę podziału binarnego minimalizującego entropię

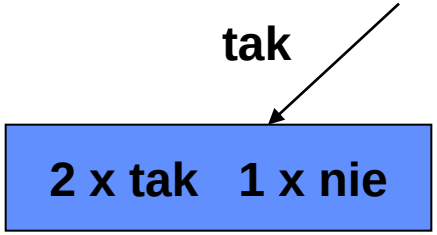
GDAŃSK UNIVERSITY OF TECHNOLOGY

Atrybuty numeryczne (2)

S.C.	D.O.R.	Wiek	Wykształcenie	Sam.	Z.K.
S	800	32	wyższe	tak	tak
S	1200	35	średnie	tak	tak
S	700	26	podstawowe	nie	nie
M	600	45	wyższe	nie	tak
M	650	38	średnie	tak	tak
S	900	28	wyższe	nie	nie
S	1100	65	średnie	tak	nie
M	500	22	średnie	nie	nie
S	800	43	podstawowe	tak	nie

1. **W trakcie budowy drzewa doszliśmy do wydzielenia 3 przykładów**
2. **W tym miejscu drzewa najlepiej zastosować podział względem wartości atrybutu numerycznego (wcześniej oczywiście takie podziały też były rozważane ale odrzucane)**

Dalsza budowa drzewa (2)



$$I(P) \approx 0,92$$

Wiek:	**35**	**38**	**65**
Z.K.:	**tak**	**tak**	**nie**

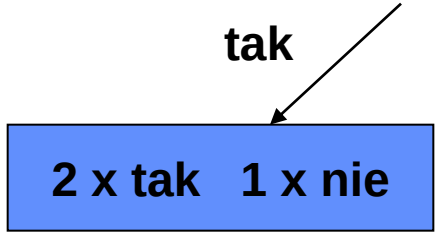
$$t = "Wiek <= x"$$
$$R_t = \{<= x, > x\}$$

$$I(P_{Wiek<=x}) \approx 0 \qquad I(P_{Wiek>x}) \approx 1$$

$$E_t(P) \approx \frac{1}{3} \cdot 0 + \frac{2}{3} \cdot 1 \approx 0,67$$

$$g_t(P) \approx 0,92 - 0,67 \approx 0,25$$

Dalsza budowa drzewa (3)



$$I(P) \approx 0,92$$

Wiek:	**35**	**38**	**65**
Z.K.:	**tak**	**tak**	**nie**

$$t = "Wiek <= x"$$
$$R_t = \{<= x, > x\}$$

$$I(P_{Wiek<=x}) \approx 0 \qquad I(P_{Wiek>x}) \approx 0$$

$$E_t(P) \approx \frac{1}{3} \cdot 0 + \frac{2}{3} \cdot 0 = 0$$

$$g_t(P) \approx 0,92 - 0 \approx 0,92$$

GDAŃSK UNIVERSITY OF TECHNOLOGY

Algorytm budowy drzew decyzyjnych

- Budowa drzewa polega na doborze najlepszego atrybutu nominalnego lub najlepszego podziału binarnego atrybutu numerycznego, powtarzanym iteracyjnie,

- Rozszerzenia:
  - Obsługa brakujących wartości atrybutów,
  - Przycinanie drzew – generalizacja.

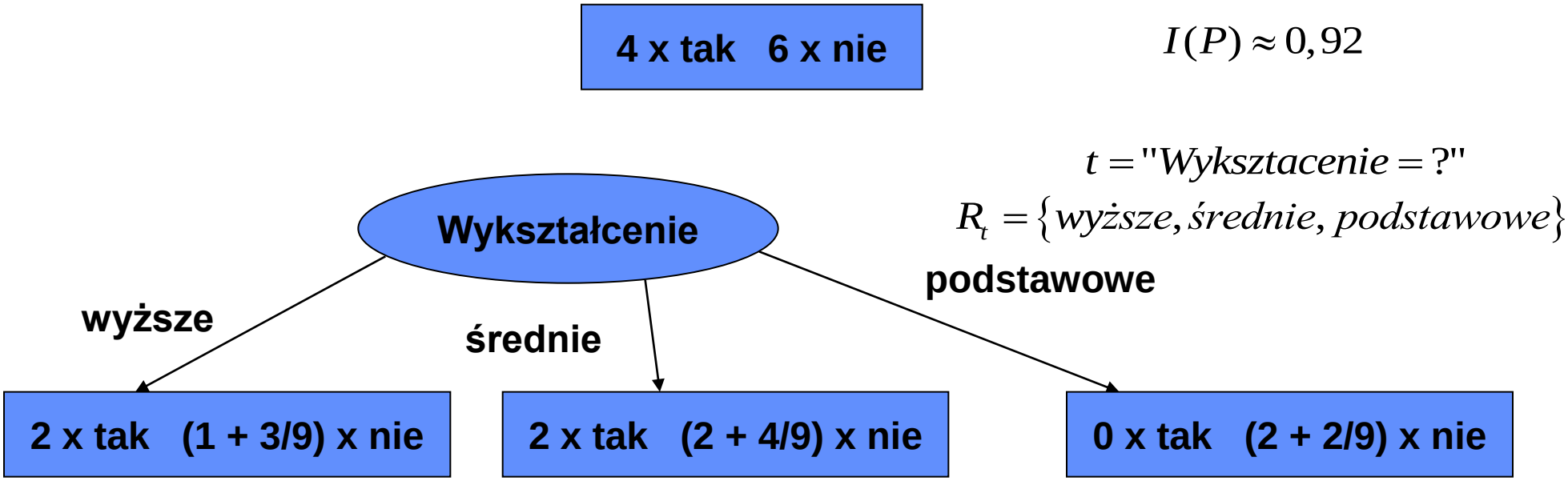
GDAŃSK UNIVERSITY OF TECHNOLOGY

Brakujące wartości atrybutów

S.C.	D.O.R.	Wiek	Wykształcenie	Sam.	Z.K.
S	800	32	wyższe	tak	tak
S	1200	35	średnie	tak	tak
S	700	26	podstawowe	nie	nie
M	600	45	wyższe	nie	tak
M	650	38	średnie	tak	tak
S	900	28	wyższe	nie	nie
S	1100	65	średnie	tak	nie
M	500	22	średnie	nie	nie
S	800	43	podstawowe	tak	nie
M	750	47	?	nie	nie

Zakładamy, że mamy dodatkowy przykład o nieznanej wartości atrybutu *Wykształcenie*

Brakujące wartości atrybutów (2)



$$I(P) \approx 0,92$$

$$t = "Wykształcenie = ?"$$

$$R_t = \{wyższe, średnie, podstawowe\}$$

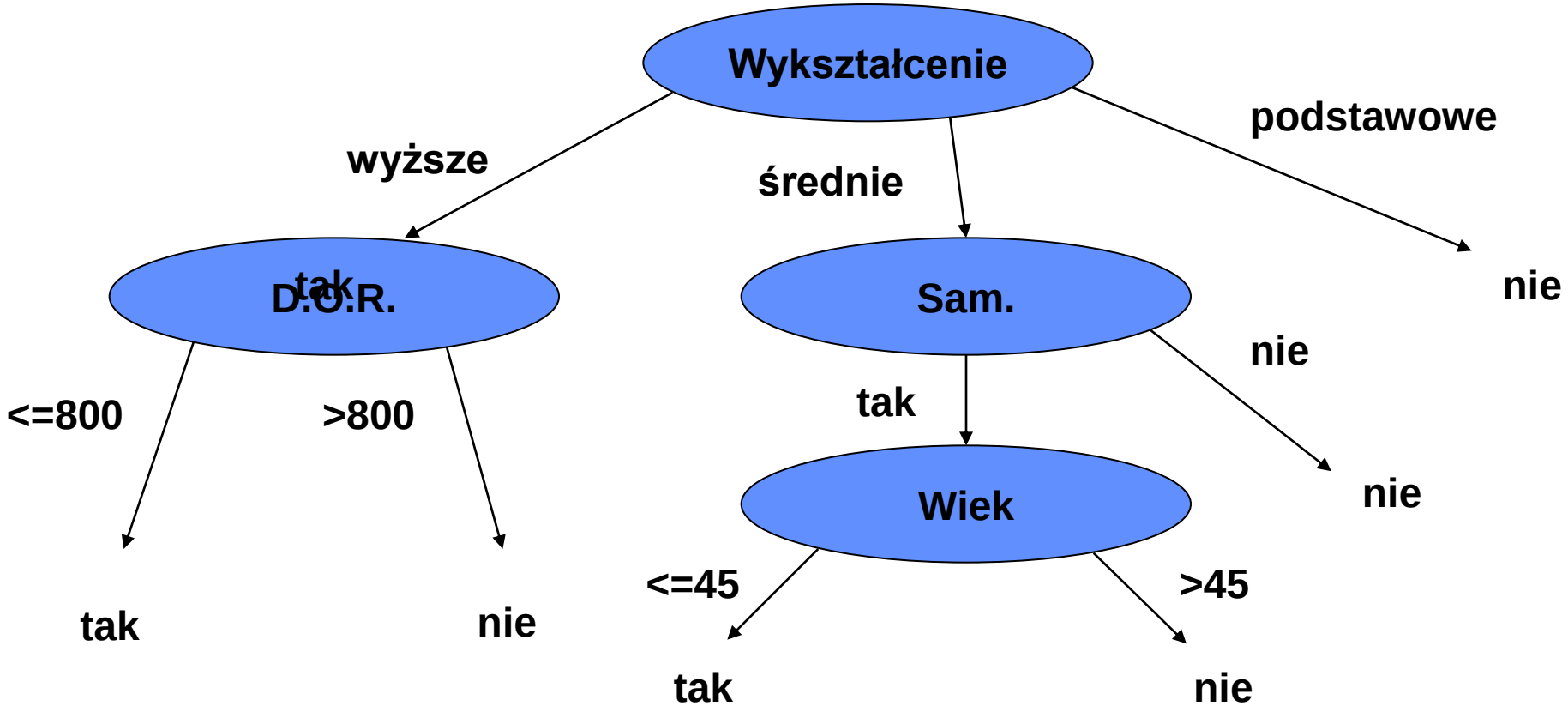
$$I(P_{Wykształcenie=wyższe}) \approx 0,97 \qquad I(P_{Wykształcenie=średnie}) = 0,99 \qquad I(P_{Wykształcenie=podstawowe}) = 0$$

$$E_t(P) \approx \frac{3+3/9}{10} \cdot 0,97 + \frac{4+4/9}{10} \cdot 0,99 + \frac{2+2/9}{10} \cdot 0 \approx 0,76$$

$$g_t(P) \approx 0,92 - 0,76 \approx 0,16$$

GDAŃSK UNIVERSITY OF TECHNOLOGY

Przycinanie drzewa decyzyjnego



GDAŃSK UNIVERSITY OF TECHNOLOGY

Przycinanie

- Przycinanie polega na zastąpieniu poddrzewa liściem,
- Przycinanie ma na celu uogólnienie wyników i zapobieżenie błędowi nadmiernego dopasowania,
- Stosuje się różne kryteria przycinania:
  - przycinanie *apriori* (w trakcie pracy zasadniczego algorytmu), gdy węzeł drzewa pokrywa zbyt małą liczbę przykładów,
  - przycinanie *aposteriori* (po pracy zasadniczego algorytmu), najczęściej *wstępująca* w wyniku badania rezultatów klasyfikacji na *zbiorze testującym*,
- W wyniku przycinania liście stają się *węzłami probabilistycznymi*

Redukcjonistyczne podejście do opisu algorytmów

Algorytm budowy drzew decyzyjnych (~C4.5):

1. Zadanie: predykcja (klasyfikacja)
2. Struktura modelu: drzewo
3. Funkcja oceny jakości: przyrost zawartości informacyjnej
4. Metody przeszukiwania: zachłanna, divide-and-conquer
5. Dodatkowe założenia:

   Obsługa brakujących wartości atrybutów metodą podziału przykładu

   Obsługa atrybutów numerycznych metodą podziału binarnego minimalizującego entropię

   Przycinanie drzewa metodą wstępującą aposteriori (walidacja krzyżowa)

GDAŃSK UNIVERSITY OF TECHNOLOGY

Dziękujemy za uwagę

Zapraszamy na wykład:

KLASYFIKACJA I REGRESJA cz. 2