Klasyfikacja

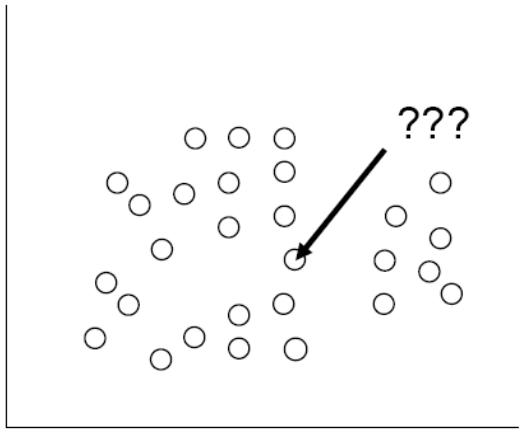
w wykładzie wykorzystano:

1. materiały dydaktyczne przygotowane w ramach projektu *Opracowanie programów nauczania na odległość na kierunku studiów wyższych – Informatyka*
   http://wazniak.mimuw.edu.pl
2. *Internetowy Podręcznik Statystyki*
   http://www.statsoft.pl/textbook/stathome.html

Krzysztof Regulski,
WIMiIP, KISiM,
regulski@agh.edu.pl
B5, pok. 408

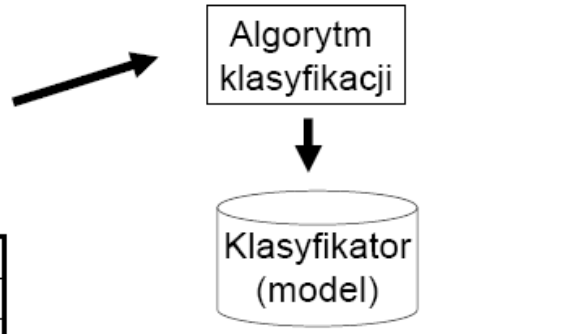
Klasyfikacja

- Klasyfikacja jest metodą analizy danych, której celem jest predykcja wartości określonego atrybutu w oparciu o pewien zbiór danych treningowych.
- Obejmuje metody odkrywania modeli (tak zwanych klasyfikatorów) lub funkcji opisujących zależności pomiędzy charakterystyką obiektów a ich zadaną klasyfikacją.
- Odkryte modele klasyfikacji są wykorzystywane do klasyfikacji nowych obiektów o nieznanej klasyfikacji.



Wiele technik:
- statystyka,
- drzewa decyzyjne,
- sieci neuronowe, etc.

Klasyfikacja – algorytm



dane treningowe

Wiek	Typ_sam	Ryzyko
20	Combi	High
18	Sports	High
40	Sports	High
50	Family	Low
35	Minivan	Low
30	Combi	High
32	Family	Low
40	Combi	Low

if Wiek < 31
or Typ_sam = „Sports"
then Ryzyko = High

Atrybut *Ryzyko* związany z informacją, że dany kierowca spowodował wcześniej wypadki czy nie powodował wcześniej wypadku. Jeżeli jest autorem kilku wypadków wartość atrybutu *Ryzyko* przyjmuje wartość High, w przypadku gdy nie spowodował żadnego wypadku atrybut Ryzyko przyjmuje wartość Low. Atrybut Ryzyko jest atrybutem decyzyjnym.

W naszym przykładzie wynikiem działania algorytmu klasyfikacji jest klasyfikator w postaci pojedynczej reguły decyzyjnej: *„Jeżeli wiek kierowcy jest mniejszy niż 31 lub typ samochodu sportowy to Ryzyko jest wysokie"*.

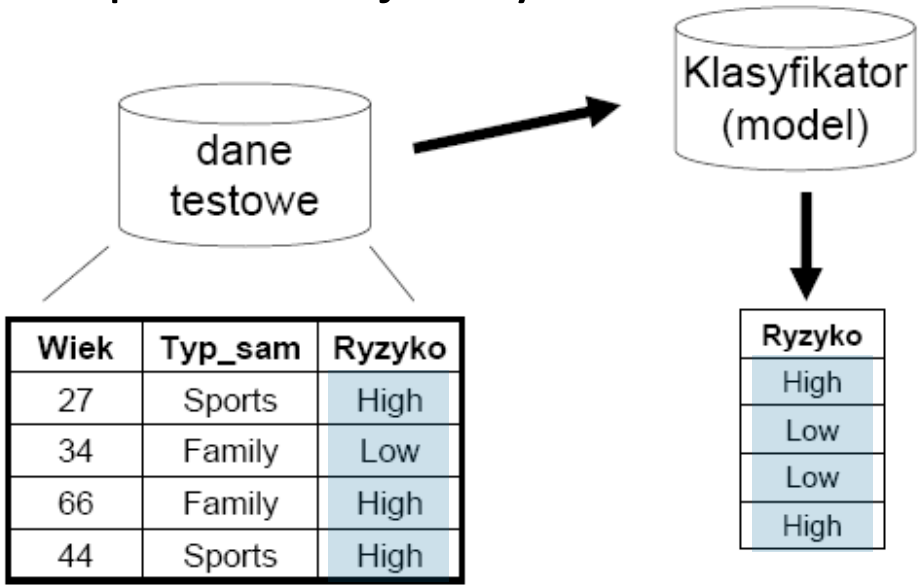
Klasyfikacja – wynik

- Wynik klasyfikacji:
  - » Reguły klasyfikacyjne postaci *IF - THEN*
  - » Formuły logiczne
  - » Drzewa decyzyjne
- Istotną sprawą z punktu widzenia poprawności i efektywności modelu jest tzw. dokładność modelu.
- dla przykładów testowych, dla których znane są wartości atrybutu decyzyjnego, wartości te są porównywane z wartościami atrybutu decyzyjnego generowanymi dla tych przykładów przez klasyfikator.
- Miarą, która weryfikuje poprawność modelu jest współczynnik dokładności.

Współczynnik dokładności (*ang. accuracy rate*) = %
procent przykładów testowych
poprawnie zaklasyfikowanych przez model

Klasyfikacja – wynik

Jeżeli dokładność klasyfikatora jest akceptowalna, wówczas możemy wykorzystać klasyfikator do klasyfikacji nowych danych.
Celem klasyfikacji, jak pamiętamy jest przyporządkowanie nowych danych dla których wartość atrybutu decyzyjnego nie jest znana do odpowiedniej klasy.



Wiek	Typ_sam	Ryzyko
27	Sports	High
34	Family	Low
66	Family	High
44	Sports	High

Ryzyko
High
Low
Low
High

Dokładność = 3/4 = 75%

Rodzaje modeli klasyfikacyjnych:

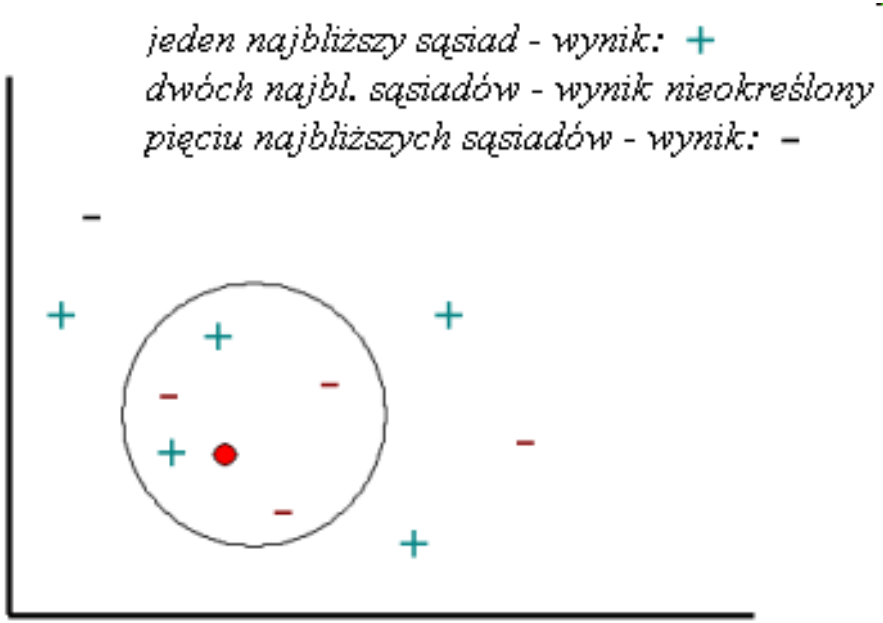
- k-NN - k-najbliższe sąsiedztwo
- Klasyfikatory Bayes'owskie
- **Klasyfikacja poprzez indukcję drzew decyzyjnych**
- Sieci Neuronowe
- Analiza statystyczna
- Metaheurystyki (np. algorytmy genetyczne)
- SVM – *Support Vector Machine* – Metoda wektorów nośnych
- Zbiory przybliżone

Klasyfikatory kNN

1. Klasyfikator *kNN - klasyfikator k-najbliższych sąsiadów* (*k-nearest neighbor classifier)*
2. Klasyfikacja nowych przypadków jest realizowana „na bieżąco", tj. wtedy, gdy pojawia się potrzeba.
3. Należy do grupy algorytmów opartych o analizę przypadku. Algorytmy te prezentują swoją wiedzę o świecie w postaci zbioru przypadków lub doświadczeń.
4. Idea klasyfikacji polega na metodach wyszukiwania tych zgromadzonych przypadków, które mogą być zastosowane do klasyfikacji nowych sytuacji.

kNN - przykład

Mamy przypadki dodatnie i ujemne oraz nowy punkt, oznaczony na czerwono. Zadanie polega na zaklasyfikowaniu nowego obiektu do plusów lub minusów, na bazie tego, z kim sąsiaduje.

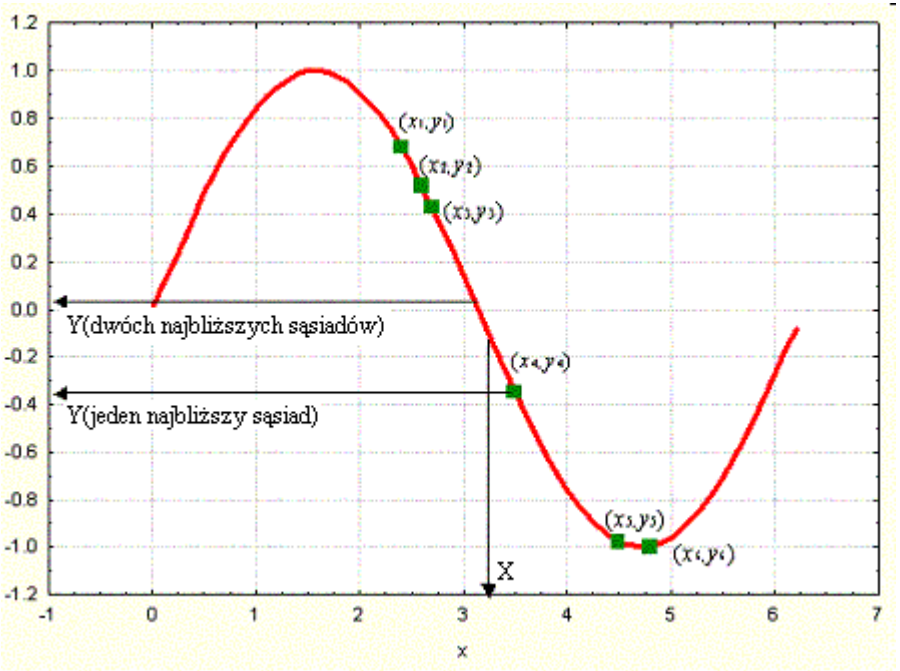


1. Zacznijmy od rozpatrzenia przypadku jednego, najbliższego sąsiada. Widać, że najbliżej czerwonego punktu jest plus, tak więc nowy przypadek zostanie zaklasyfikowany do plusów.

2. Zwiększmy teraz liczbę najbliższych sąsiadów do dwóch. Niestety, jest kłopot, drugi sąsiad to minus, więc plusy i minusy występują w tej samej ilości, nikt nie ma przewagi.

3. Zwiększmy dalej liczbę najbliższych sąsiadów, do pięciu. Są to przypadki znajdujące się wewnątrz kółka na rysunku. Jest tam przewaga minusów, więc nowy przypadek oceniamy jako minus.

Regresja i kNN



Mamy danych kilka "przykładowych" punktów, a podać musimy wartość *y* dla dowolnego *x*.

dla pojedynczego najbliższego sąsiada: Najbliżej nowego *X* znajduje się punkt o odciętej x_4. Tak więc, jako wartość dla nowego *X* przyjęta będzie wartość (rzędna) odpowiadająca x_4, czyli y_4. Oznacza to, że dla jednego najbliższego sąsiada wynikiem jest

$Y = y_4$

dwóch najbliższych sąsiadów: szukamy dwóch punktów mających najbliżej do *X*.
Są to punkty o wartościach rzędnych y_3 i y_4. Biorąc średnią z dwóch wartości, otrzymujemy:

$$Y = \frac{y_3 + y_4}{2}$$

W podobny spos my przy dowolnej liczbie *K* najbliższych sąsiadów. Wartość Y zmiennej zależnej otrzymujemy jako średnią z wartości zmiennej zależnej dla *K* punktów o wartościach zmiennych niezależnych X najbliższych *nowemu* X-owi.

Klasyfikacja w oparciu o Naiwny klasyfikator Bayesa

- Zadaniem klasyfikatora Bayes'a jest przyporządkowanie nowego przypadku do jednej z klas decyzyjnych, przy czym zbiór klas decyzyjnych musi być skończony i zdefiniowany *a priori*.
- Naiwny klasyfikator Bayes'a jest statystycznym klasyfikatorem, opartym na twierdzeniu Bayesa.
- *P(C|X)* prawdopodobieństwo *a posteriori*, że przykład *X* należy do klasy *C*
- Naiwny klasyfikator Bayes'a różni się od zwykłego klasyfikatora tym, że konstruując go zakładamy wzajemną niezależność atrybutów opisujących każdy przykład.

Przykład X klasyfikujemy jako pochodzący z tej klasy C_i, dla której wartość $P(C_i|X)$, i = 1, 2, ..., m, jest największa

Naiwny klasyfikator Bayesa

W jaki sposób oszacować prawdopodobieństwo a-posteriori P(C|X)?

Twierdzenie Bayesa:

$$P(C|X) = (P(X|C) * P(C))/P(X),$$

P(C) oznacza prawdopodobieństwo a-priori wystąpienia klasy C (tj. prawdopodobieństwo, że dowolny przykład należy do klasy C),
P(X|C) oznacza prawdopodobieństwo a-posteriori, że X należy do klasy C,
P(X) oznacza prawdopodobieństwo a-priori wystąpienia przykładu X

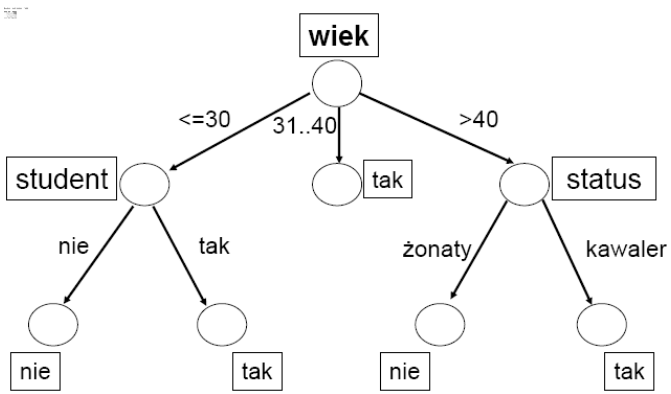
Przykład:
Dany zbiór przykładów opisujących wnioski kredytowe klientów banku

P(Ryzyko=niskie | Wiek=38, Status=rozwodnik, Dochód=niski, Dzieci=2)

oznacza prawdopodobieństwo a-posteriori, że klient, X=(38, rozwodnik, niski, 2), składający wniosek kredytowy jest klientem o niskim ryzyku kredytowym (klient wiarygodny)

Przykład

Chcemy dokonać predykcji klasy, do której należy nowy przypadek
C_1 (kupi_komputer ='tak')
C_2 (kupi_komputer ='nie')
Nowy przypadek:
X = (wiek='<=30', dochód='średni', student = 'tak', status='kawaler')
Maksymalizujemy wartość
$P(X/C_i)*P(C_i)$, dla $i=1,2$



P(kupi_komputer = 'tak') = P(C1) = 9/14 = 0.643
P(kupi_ komputer = 'nie') = P(C2) = 5/14 = 0.357

P(wiek <= '30' | kupi_ komputer = 'tak')	= 2/9	= 0.222
P(wiek <= '30' | kupi_ komputer = 'nie')	= 3/5	= 0.6
P(dochód = 'średni' | kupi_ komputer = 'tak')	= 4/9	= 0.444
P(dochód = 'średni' | kupi_ komputer = 'nie')	= 2/5	= 0.4
P(student = 'tak'| kupi_ komputer = 'tak')	= 6/9	= 0.667
P(student = 'tak'| kupi_ komputer = 'nie')	= 1/5	= 0.2
P(status = 'kawaler'| kupi_ komputer = 'tak')	= 6/9	= 0.667
P(status ='kawaler'| kupi_ komputer = 'nie')	= 2/9	= 0.4

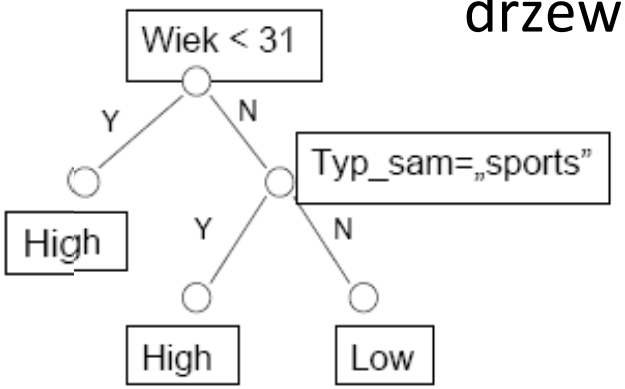
Korzystając z obliczonych prawdopodobieństw, otrzymujemy:

P(X| kupi_komputer='tak') = 0.222 * 0.444 * 0.667 * 0.667 = 0.044
P(X| kupi_komputer='nie') = 0.600 * 0.400 * 0.200 * 0.400 = 0.019
Stąd:
P(X| kupi_ komputer='tak') * P(kupi_ komputer='tak') = 0.044 * 0.643 = **0.028**
P(X| kupi_ komputer='nie') * P(kupi_ komputer='nie') = 0.019 * 0.357 = 0.007

Naiwny klasyfikator Bayesa zaklasyfikuje nowy przypadek X do klasy:

kupi_komputer = 'tak'

Indukcja drzew decyzyjnych



Drzewo decyzyjne jest grafem o strukturze drzewiastej, gdzie

» każdy wierzchołek wewnętrzny reprezentuje test na atrybucie (atrybutach),

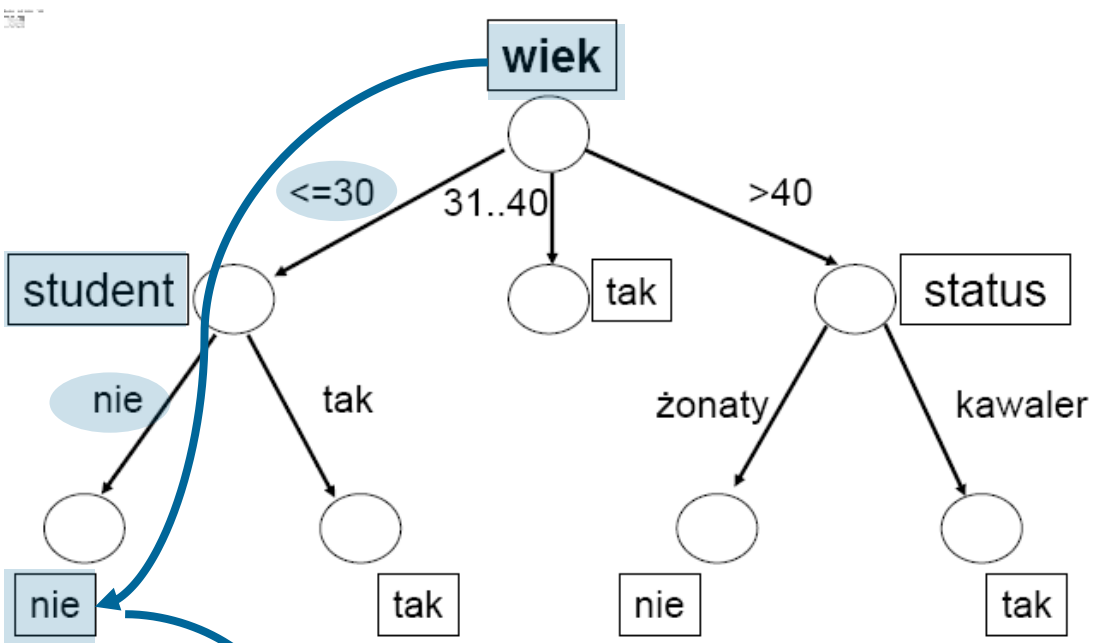
» każdy łuk reprezentuje wynik testu,

» każdy liść reprezentuje pojedynczą klasę lub rozkład wartości klas

Drzewo decyzyjne dzieli zbiór treningowy na partycje do momentu, w którym każda partycja zawiera dane należące do jednej klasy, lub, gdy w ramach partycji dominują dane należące do jednej klasy

Każdy wierzchołek wewnętrzny drzewa zawiera tzw. **punkt podziału** *(ang. split point),* którym jest test na atrybucie (atrybutach), który dzieli zbiór danych na partycje

Ekstrakcja reguł klasyfikacyjnych z drzew decyzyjnych

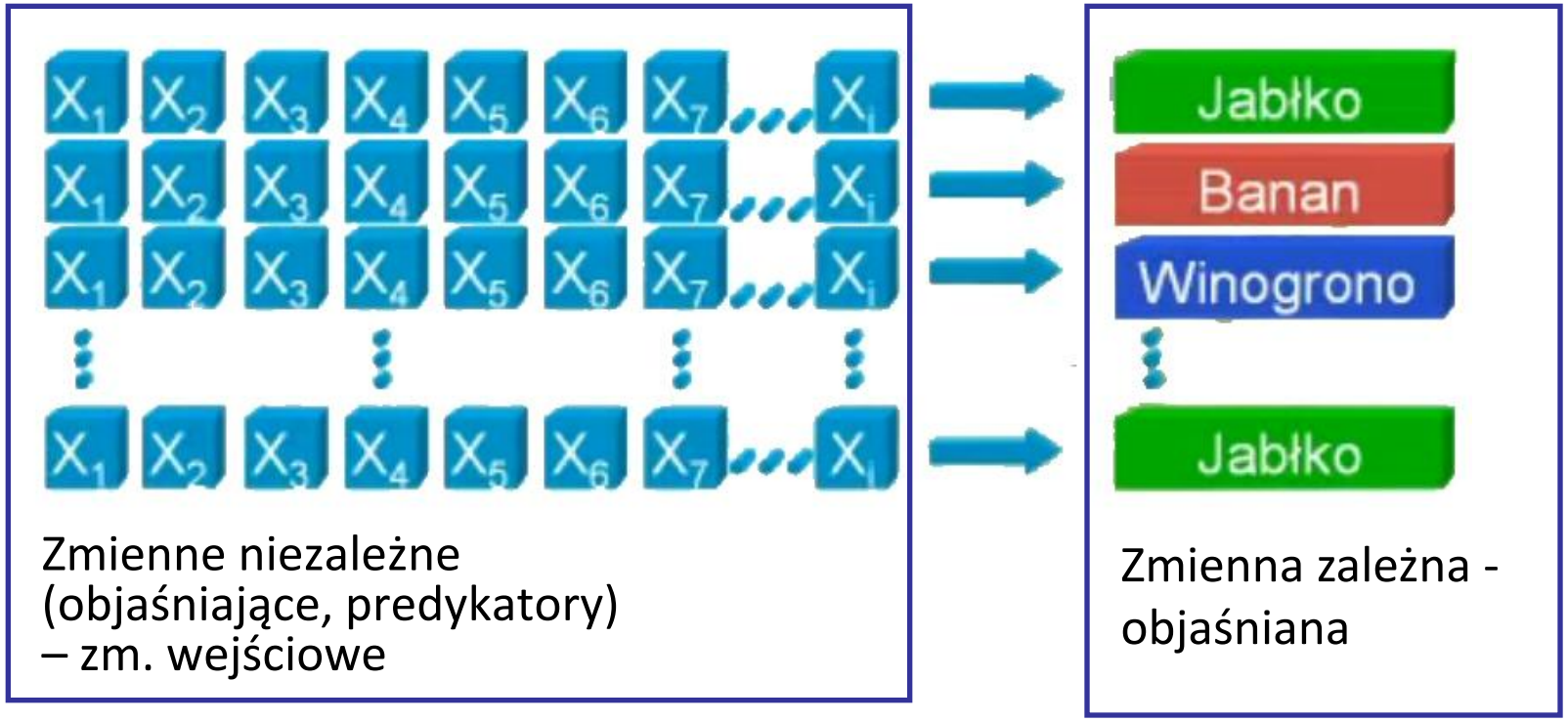


- Drzewo decyzyjne można przedstawić w postaci zbioru tzw. reguł klasyfikacyjnych postaci **IF-THEN**
- Dla każdej ścieżki drzewa decyzyjnego, łączącej korzeń drzewa z liściem drzewa tworzymy regułę klasyfikacyjną

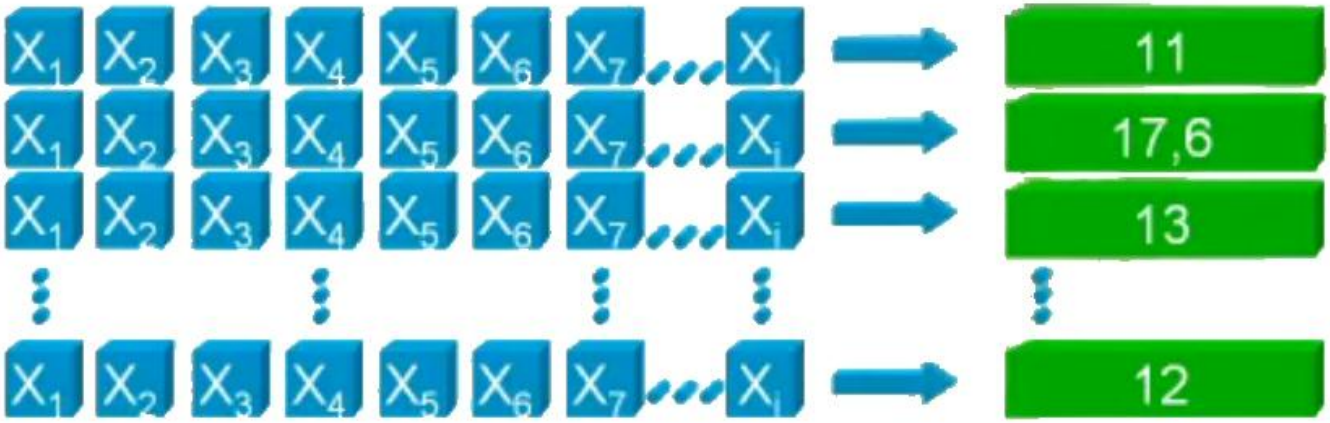
Reguły:

IF wiek='<=30' **AND** student='nie' **THEN** kupi_komputer='nie'
IF wiek ='<=30'**AND** student='tak' **THEN** kupi_komputer='tak'
IF wiek ='31..40' **THEN** kupi_komputer ='tak'
IF wiek ='>40' **AND** status='żonaty' **THEN** kupi_komputer ='nie'
IF wiek ='>40' **AND** status ='kawaler'**THEN** kupi_komputer ='tak'

Klasyfikacja / Regresja



KLASYFIKACYJNE

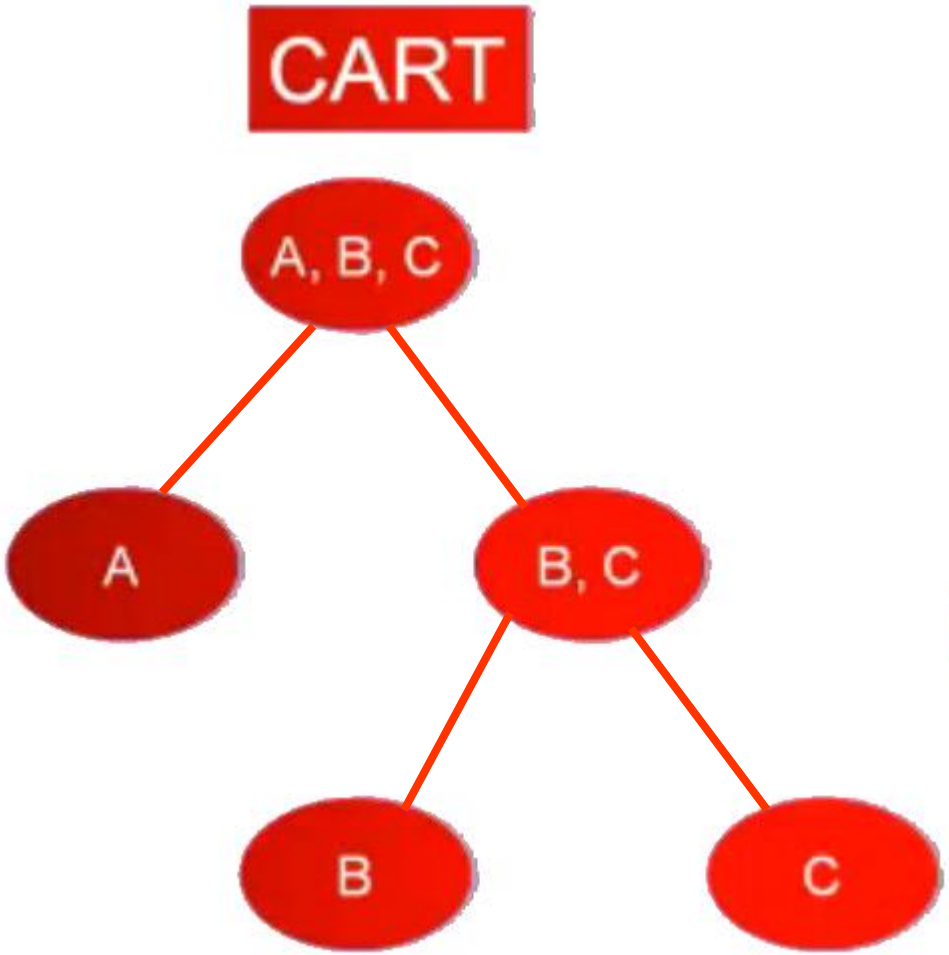


REGRESYJNE

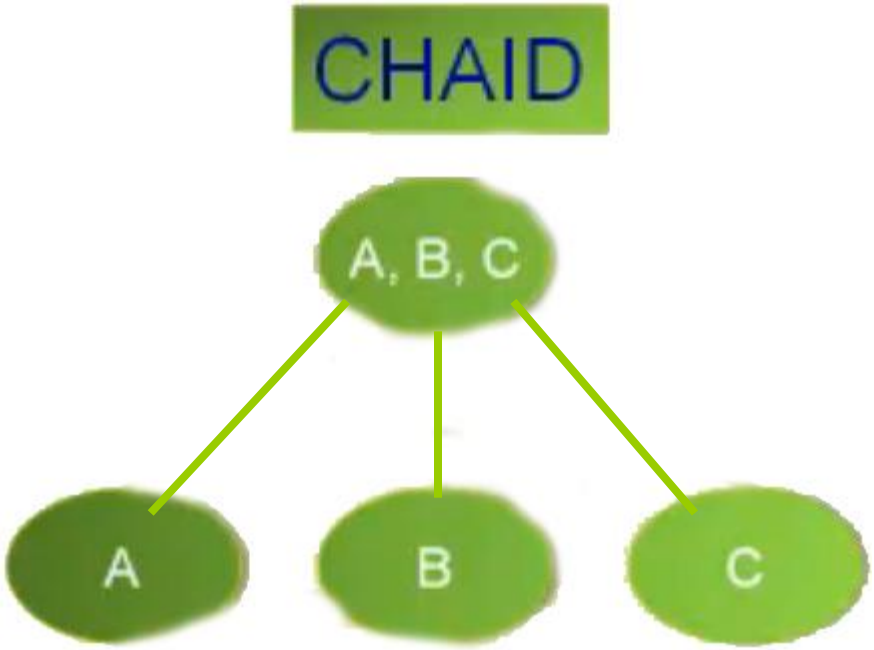
CART (C&RT) i CHAID

Classification & Regression Trees

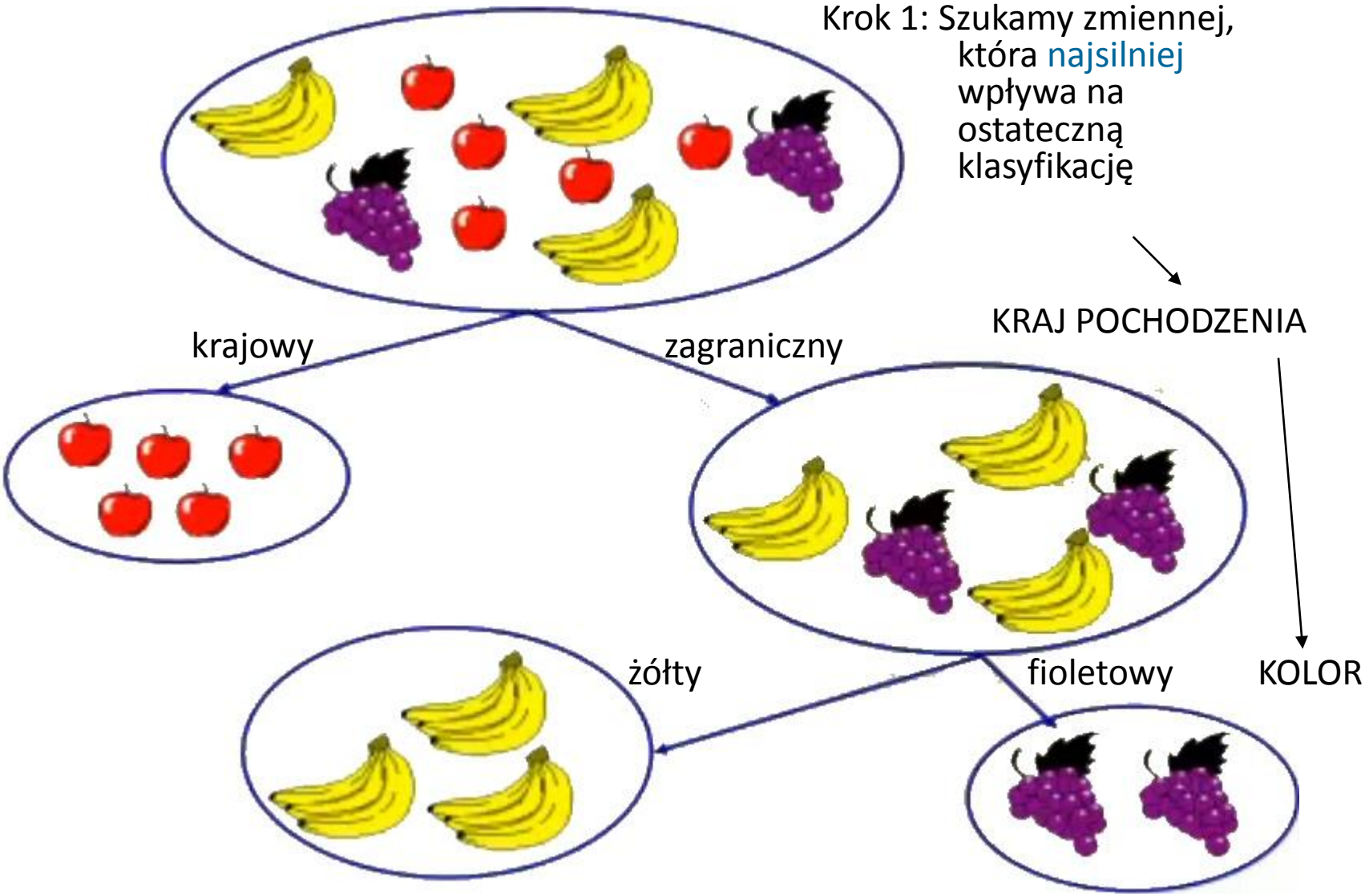
C&RT



Chi-square Automatic Interaction Detection



Drzewa klasyfikacyjne



Przykład: segmentacja pod kątem dochodów

- Cel: segmentacja ze względu na dochód, charakterystyka segmentów
- Dane zawierają 32 561 przypadków
- Każdy przypadek reprezentuje jedną osobę
- Każda osoba opisana jest przez 14 cech oraz zmienną dochód

dults (11 zmn. * 32561 prz.)

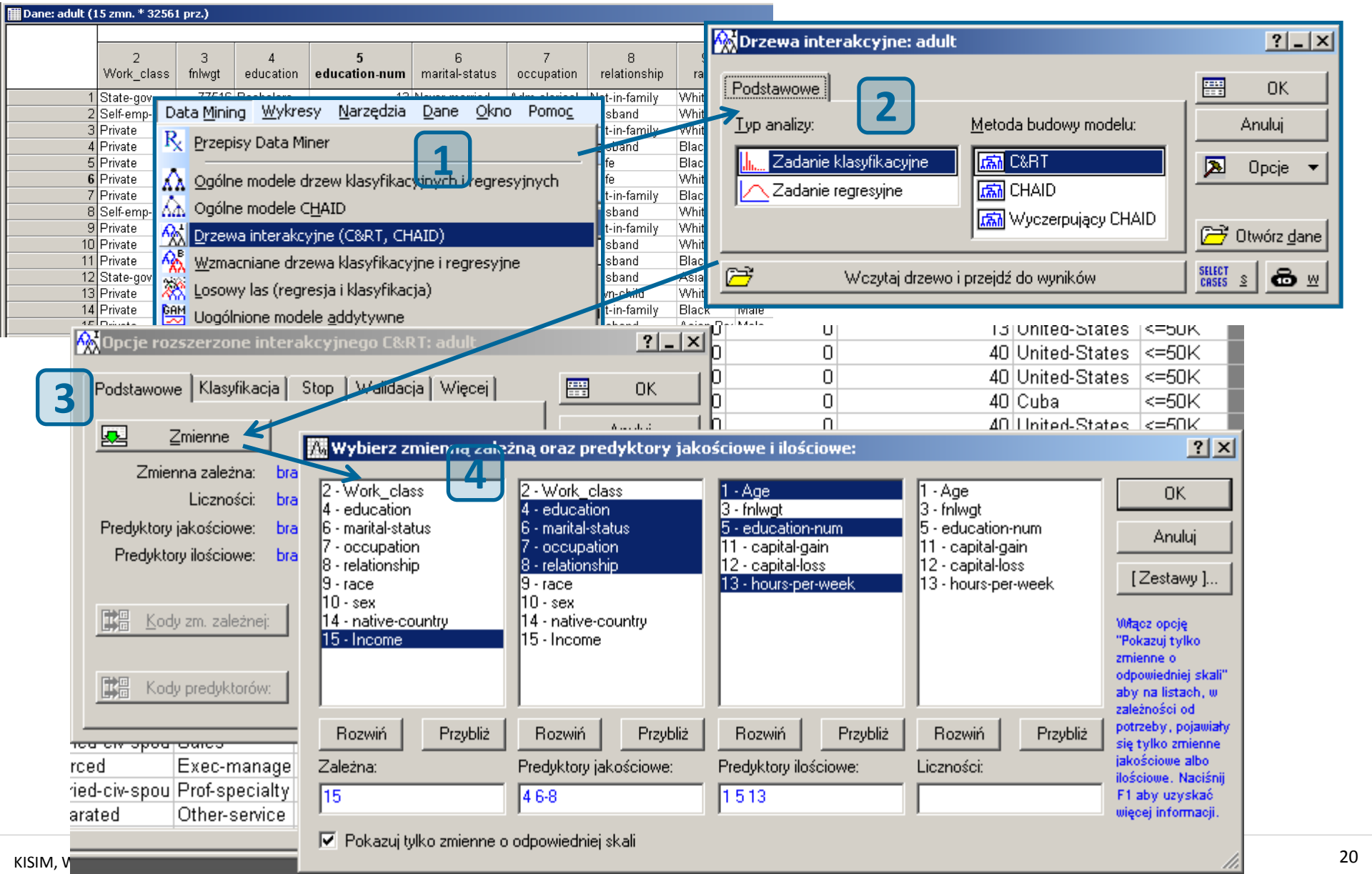
	1 Wiek	2 Grupa zawodowa	3 Wykształcenie	4 Liczba lat kształcenia	5 Stan cywilny	6 Zawód	7 Związek	8 Rasa	9 Płeć
1	25	Firma prywatna	College	10	Rozwiedziona/ny	Kierownik/Dyrektor	Stanu wolnego	Biała	Kobieta
2	39	Firma prywatna	Licencjat	12	Separacja	Mechanik	Poza rodziną	Biała	Mężczyzna
3	45	Firma prywatna	Licencjat	11	Rozwiedziona/ny	Inne usługi	Poza rodziną	Biała	Kobieta
4	29	Firma prywatna	Szkoła średnia	9	Małżeństwo	Wsparcie techniczne	Wychowująca(cy) dziecko	Pozostałe	Mężczyzna
5	36	Firma prywatna	Licencjat	13	Małżeństwo	Sprzedawca	Mąż	Biała	Mężczyzna
6	30	Firma prywatna	College	10	Małżeństwo	Transportowiec	Mąż	Pozostałe	Mężczyzna
7	33	Firma prywatna	College	10	Separacja	Wsparcie techniczne	Stanu wolnego	Biała	Kobieta
8	42	Administracja lokalna	Szkoła średnia	9	Rozwiedziona/ny	Inne usługi	Poza rodziną	Pozostałe	Mężczyzna
9	42	Samozatrudnienie firma	Licencjat	13	Małżeństwo	Sprzedawca	Mąż	Biała	Mężczyzna
10	24	Firma prywatna	Szkoła średnia	9	Stanu wolnego	Inne usługi	Poza rodziną	Biała	Kobieta
11	23	Firma prywatna	Szkoła średnia	9	Stanu wolnego	Kierownik/Dyrektor	Poza rodziną	Pozostałe	Mężczyzna
12	31	Firma prywatna	Licencjat	13	Stanu wolnego	Urzędnik	Poza rodziną	Biała	Mężczyzna
13	31	Administracja krajowa	College	10	Małżeństwo	Inspektor	Mąż	Biała	Mężczyzna
14	32	Inne	Szkoła średnia	9	Rozwiedziona/ny	?	Stanu wolnego	Pozostałe	Mężczyzna
15	43	Firma prywatna	College	10	Małżeństwo	Inne usługi	Mąż	Biała	Mężczyzna
16	35	Firma prywatna	Szkoła średnia	9	Małżeństwo	Kierownik/Dyrektor	Żona	Biała	Kobieta
17	36	Administracja federalan	College	10	Separacja	Urzędnik	Stanu wolnego	Pozostałe	Kobieta
18	26	Firma prywatna	Podstawowa	8	Stanu wolnego	Transportowiec	Wychowująca(cy) dziecko	Pozostałe	Mężczyzna
19	33	Administracja krajowa	College	10	Rozwiedziona/ny	Inne usługi	Stanu wolnego	Pozostałe	Kobieta
20	37	Firma prywatna	Licencjat	13	Stanu wolnego	Sprzedawca	Poza rodziną	Pozostałe	Kobieta
21	44	Firma prywatna	Szkoła średnia	9	Małżeństwo	Sprzątacz	Żona	Biała	Kobieta
22	19	Inne	College	10	Stanu wolnego	?	Wychowująca(cy) dziecko	Pozostałe	Kobieta
23	32	Firma prywatna	Szkoła średnia	9	Stanu wolnego	Inspektor	Wychowująca(cy) dziecko	Pozostałe	Mężczyzna
24	33	Firma prywatna	Podstawowa	8	Stanu wolnego	Usługi ochroniarskie	Wychowująca(cy) dziecko	Pozostałe	Mężczyzna

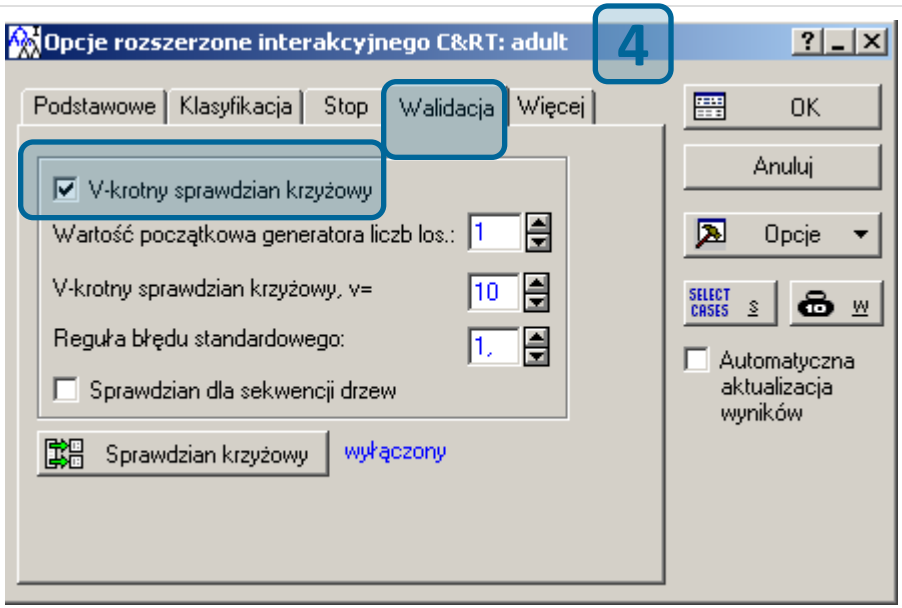
Dochód skategoryzowany do 2 klas:

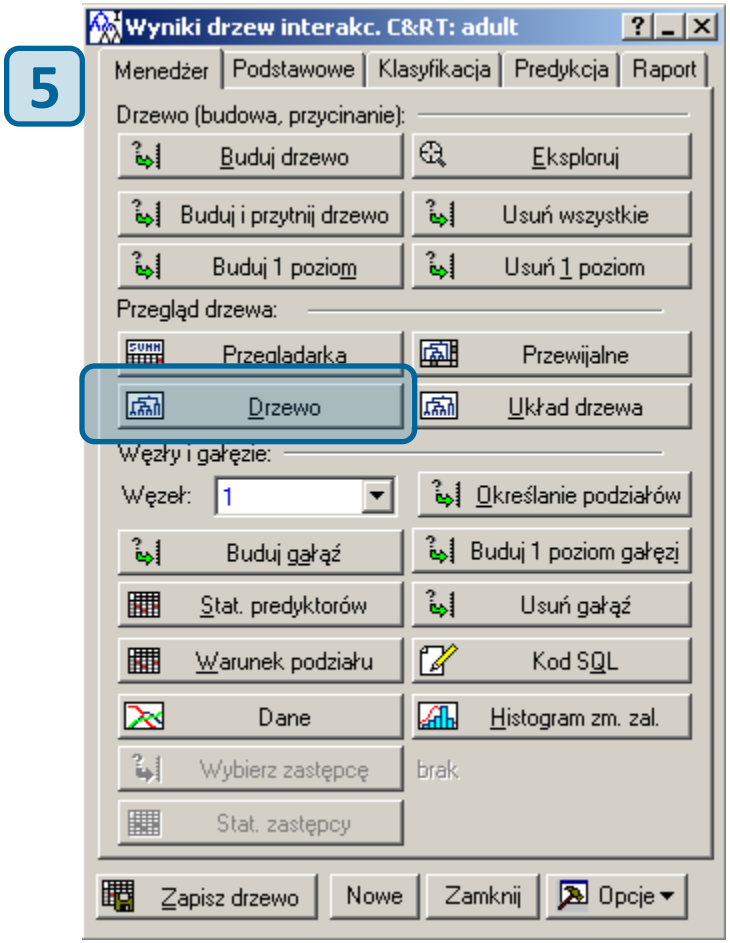
<=50K – poniżej 50 000

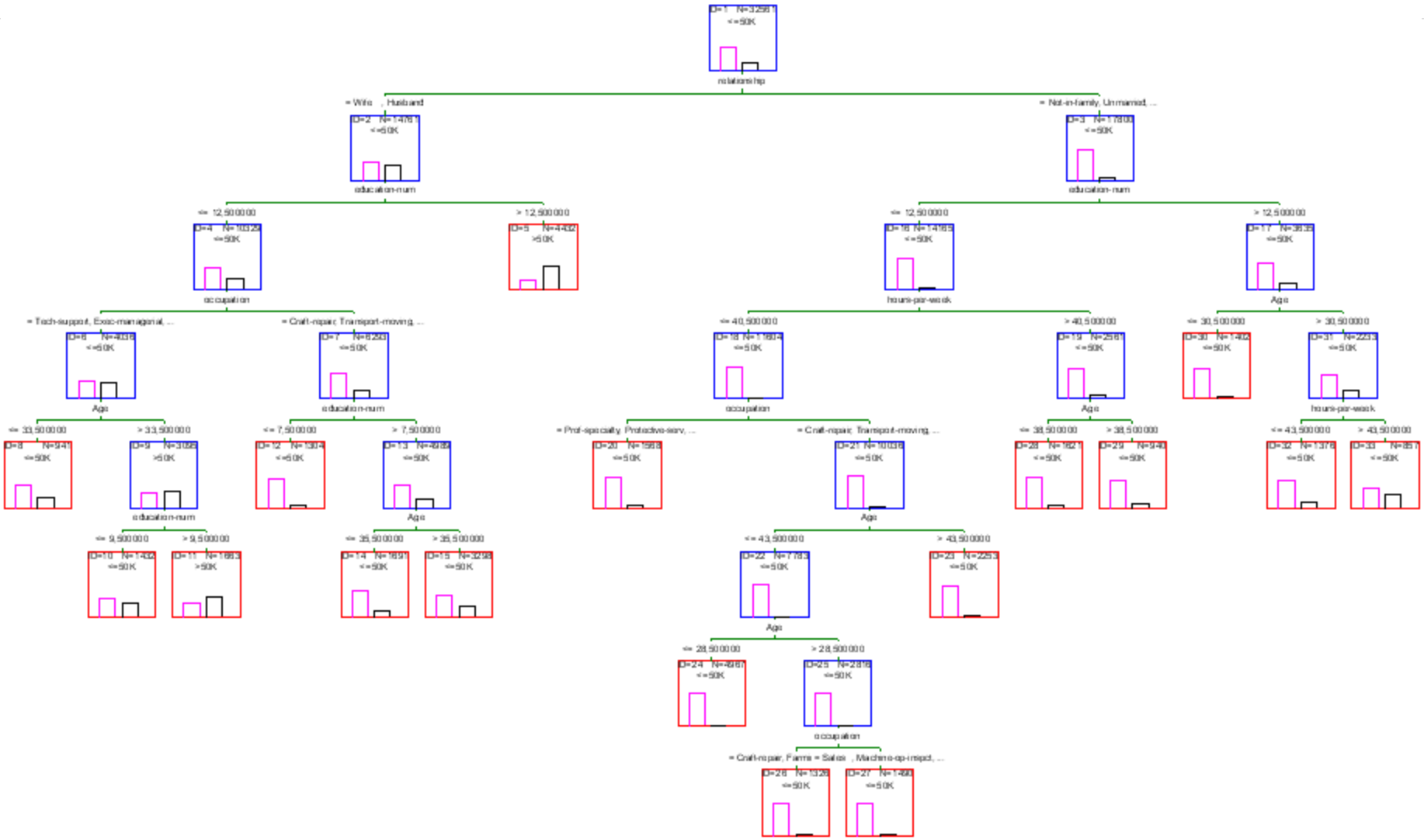
>50K – powyżej 50 000

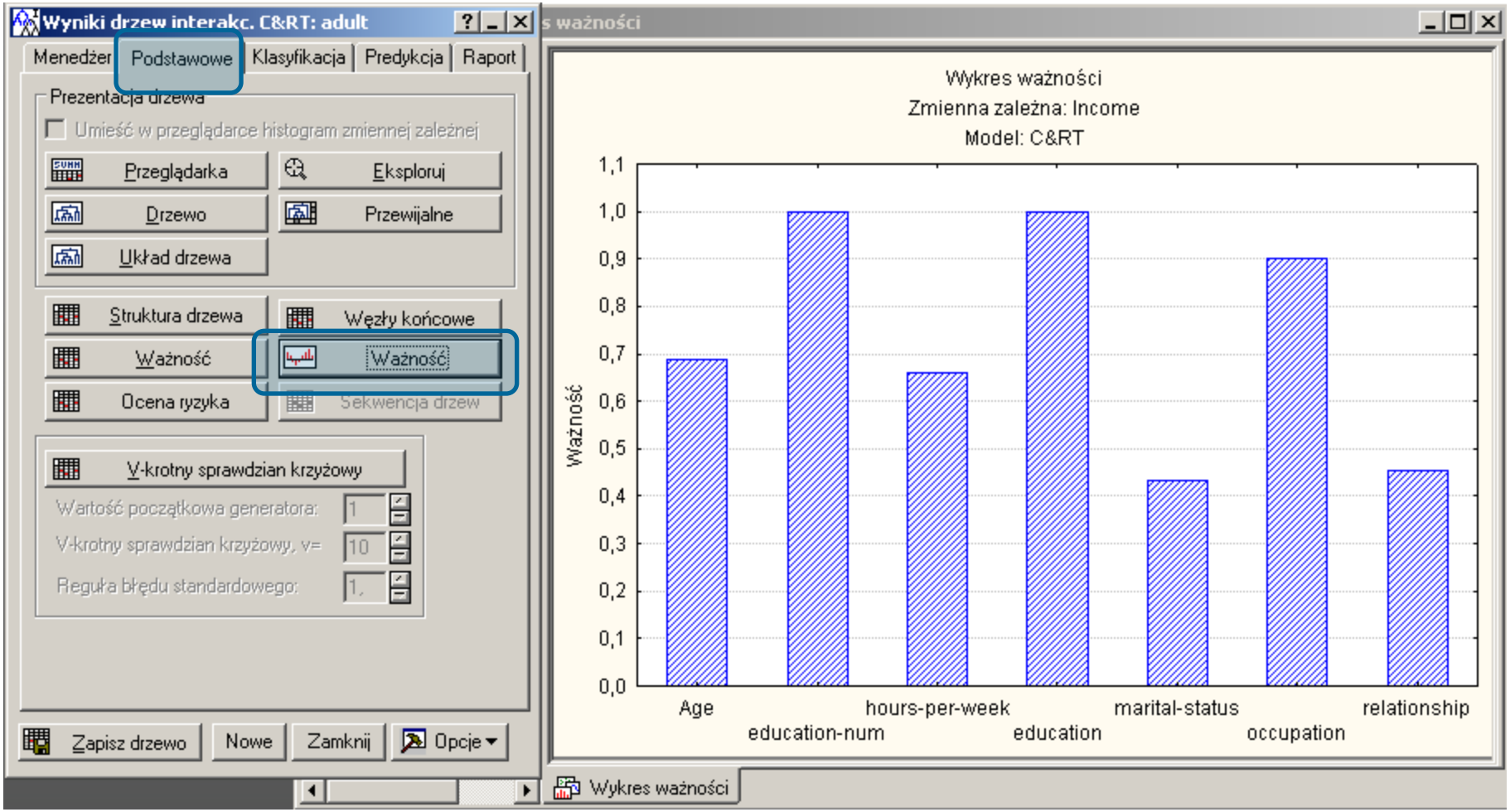
STATISTICA - przykład drzewa C&RT

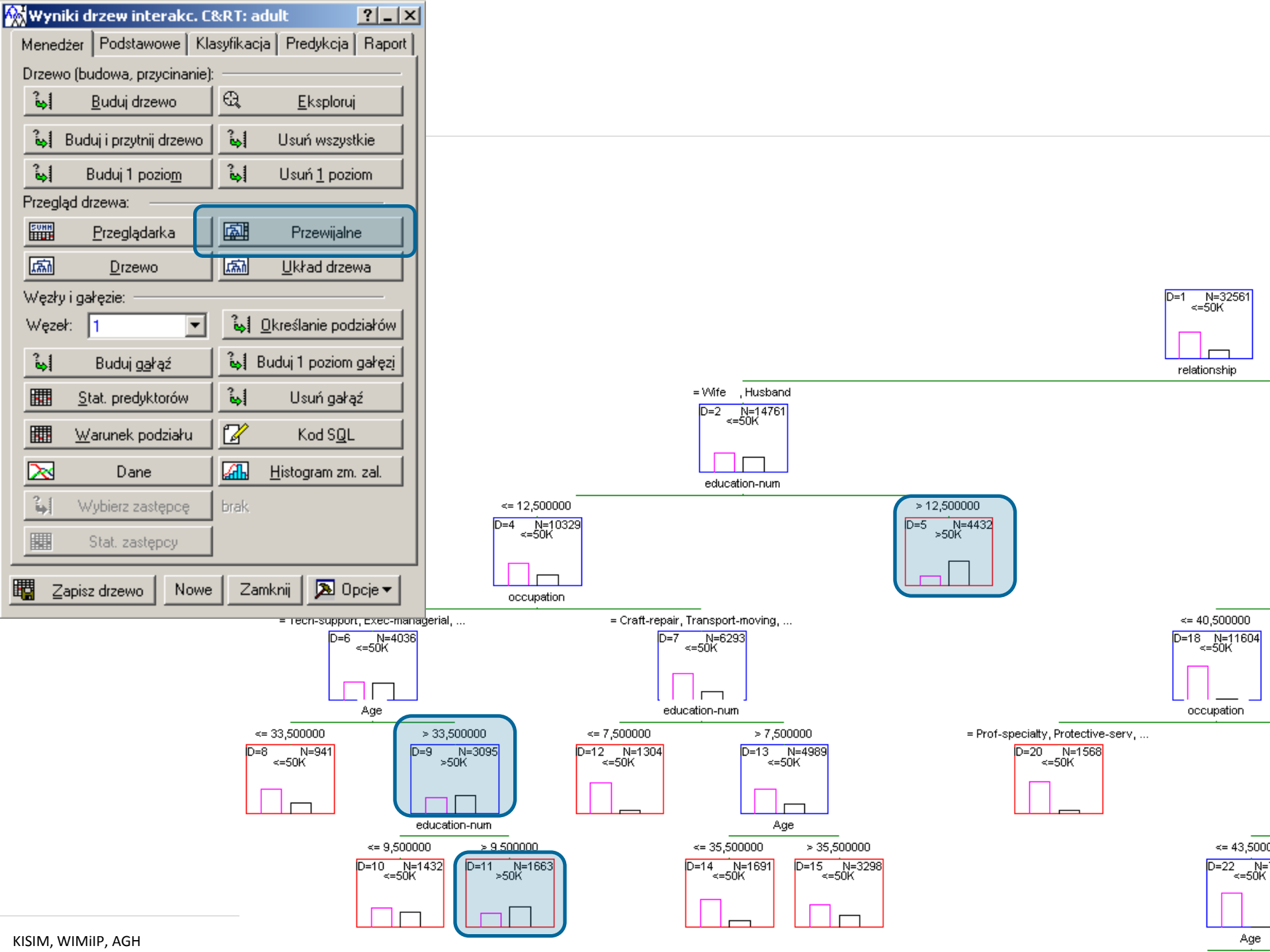


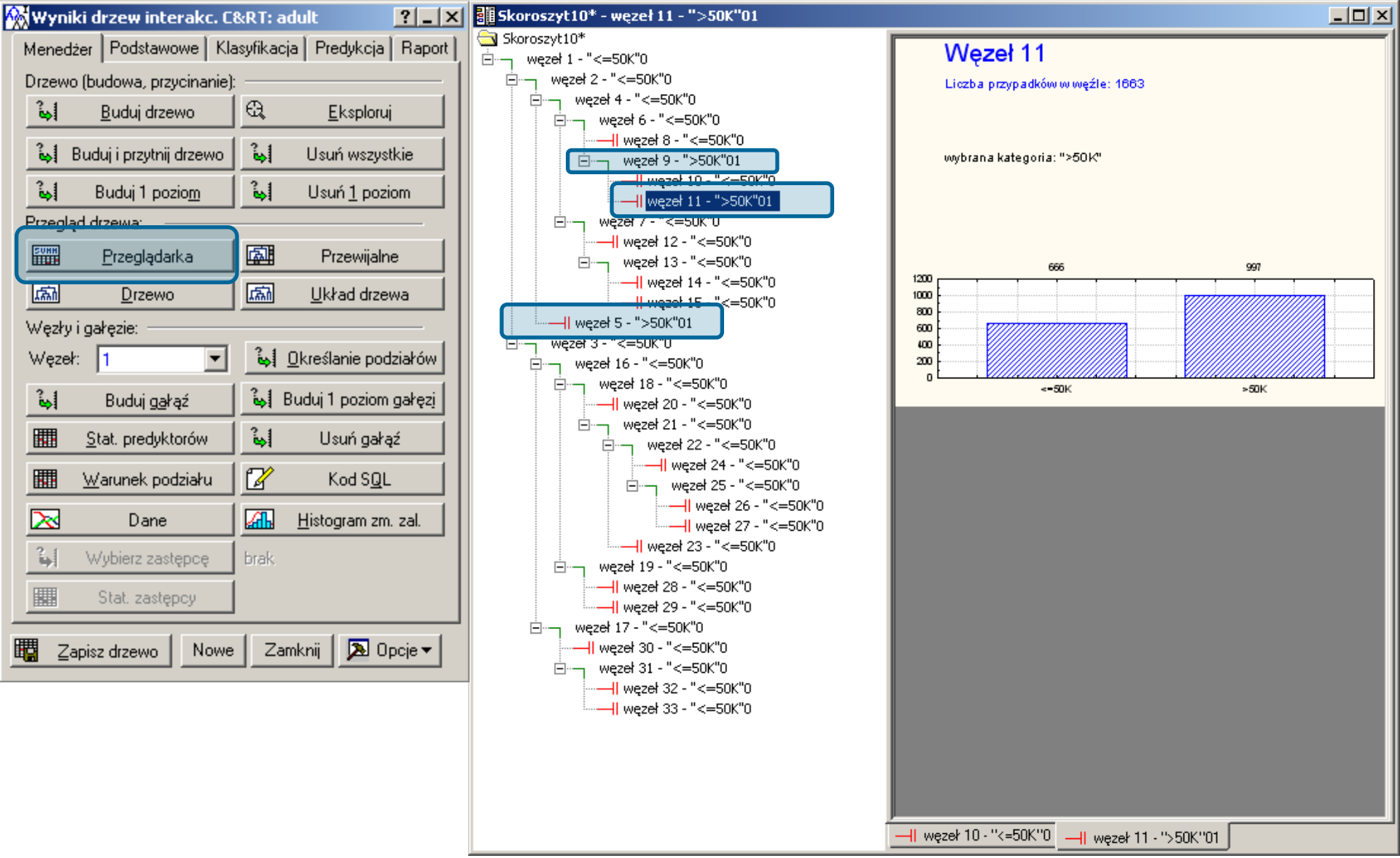

Opcje rozszerzone interakcyjnego C&RT: adult
4
Podstawowe
Klasyfikacja
Stop
Walidacja
Więcej
V-krotny sprawdzian krzyżowy
Wartość początkowa generatora liczb los.:
1
V-krotny sprawdzian krzyżowy, v=
10
Reguła błędu standardowego:
1,
Sprawdzian dla sekwencji drzew
Sprawdzian krzyżowy
wyłączony
OK
Anuluj
Opcje
SELECT CASES
s
w
Automatyczna aktualizacja wyników


Wyniki drzew interakc. C&RT: adult
5
Menedżer
Podstawowe
Klasyfikacja
Predykcja
Raport
Drzewo (budowa, przycinanie):
Buduj drzewo
Eksploruj
Buduj i przytnij drzewo
Usuń wszystkie
Buduj 1 poziom
Usuń 1 poziom
Przegląd drzewa:
Przegladarka
Przewijalne
Drzewo
Układ drzewa
Węzły i gałęzie:
Węzeł:
1
Określanie podziałów
Buduj gałąź
Buduj 1 poziom gałęzi
Stat. predyktorów
Usuń gałąź
Warunek podziału
Kod SQL
Dane
Histogram zm. zal.
Wybierz zastępcę
brak
Stat. zastępcy
Zapisz drzewo
Nowe
Zamknij
Opcje


ID=1 N=32561
<=50K
relationship
= Wife , Husband
= Not-in-family, Unmarried, ...
ID=2 N=14781
<=50K
education-num
ID=3 N=17780
<=50K
education-num
<= 12.500000
> 12.500000
ID=4 N=10525
<=50K
ID=5 N=4432
>50K
occupation
<= 12.500000
ID=16 N=14169
<=50K
hours-per-week
> 12.500000
ID=17 N=3639
<=50K
Age
= Tech-support, Exec-managerial, ...
= Craft-repair, Transport-moving, ...
ID=6 N=4316
<=50K
ID=7 N=6209
<=50K
Age
education-num
<= 40.500000
ID=18 N=11804
<=50K
occupation
> 40.500000
ID=19 N=2561
<=50K
Age
<= 30.500000
ID=30 N=1462
<=50K
> 30.500000
ID=31 N=2229
<=50K
hours-per-week
<= 33.500000
ID=8 N=941
<=50K
> 33.500000
ID=9 N=3375
>50K
<= 7.500000
ID=12 N=1304
<=50K
> 7.500000
ID=13 N=4905
<=50K
= Prof-specialty, Protective-serv, ...
ID=20 N=1568
<=50K
= Craft-repair, Transport-moving, ...
ID=21 N=10036
<=50K
<= 38.500000
ID=28 N=1621
<=50K
> 38.500000
ID=29 N=940
<=50K
<= 43.500000
ID=32 N=1378
<=50K
> 43.500000
ID=33 N=851
<=50K
education-num
Age
<= 9.500000
ID=10 N=1432
<=50K
> 9.500000
ID=11 N=1883
>50K
<= 35.500000
ID=14 N=1891
<=50K
> 35.500000
ID=15 N=3298
<=50K
<= 43.500000
ID=22 N=7783
<=50K
> 43.500000
ID=23 N=2253
<=50K
Age
<= 28.500000
ID=24 N=4961
<=50K
> 28.500000
ID=25 N=2818
<=50K
occupation
= Craft-repair, Farming-fishing, Machine-op-inspct, ...
ID=26 N=1326
<=50K
ID=27 N=1492
<=50K

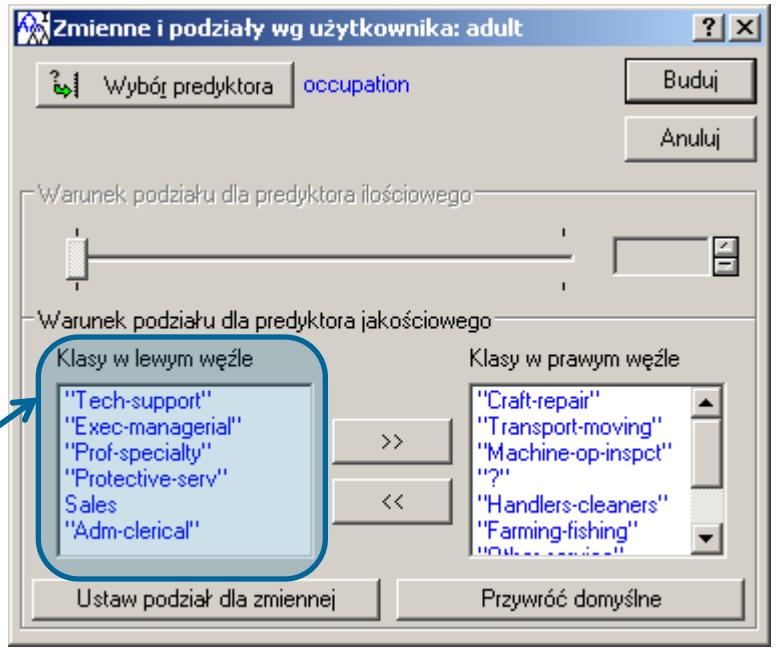

Wyniki drzew interakc. C&RT: adult
Menedżer
Podstawowe
Klasyfikacja
Predykcja
Raport
Prezentacja drzewa
Umieść w przeglądarce histogram zmiennej zależnej
Przeglądarka
Eksploruj
Drzewo
Przewijalne
Układ drzewa
Struktura drzewa
Węzły końcowe
Ważność
Ważność
Ocena ryzyka
Sekwencja drzew
V-krotny sprawdzian krzyżowy
Wartość początkowa generatora:
V-krotny sprawdzian krzyżowy, v=
Reguła błędu standardowego:
Zapisz drzewo
Nowe
Zamknij
Opcje
Wykres ważności
Wykres ważności
Zmienna zależna: Income
Model: C&RT
Ważność
1,1
1,0
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0,0
Age
education-num
hours-per-week
education
marital-status
occupation
relationship
Wykres ważności


Wyniki drzew interakc. C&RT: adult
Menedżer
Podstawowe
Klasyfikacja
Predykcja
Raport
Drzewo (budowa, przycinanie):
Buduj drzewo
Eksploruj
Buduj i przytnij drzewo
Usuń wszystkie
Buduj 1 poziom
Usuń 1 poziom
Przegląd drzewa:
Przeglądarka
Przewijalne
Drzewo
Układ drzewa
Węzły i gałęzie:
Węzeł:
1
Określanie podziałów
Buduj gałąź
Buduj 1 poziom gałęzi
Stat. predyktorów
Usuń gałąź
Warunek podziału
Kod SQL
Dane
Histogram zm. zal.
Wybierz zastępcę
brak
Stat. zastępcy
Zapisz drzewo
Nowe
Zamknij
Opcje
D=1 N=32561 <=50K
relationship
= Wife , Husband
D=2 N=14761 <=50K
education-num
<= 12,500000
D=4 N=10329 <=50K
occupation
> 12,500000
D=5 N=4432 >50K
= Tech-support, Exec-managerial, ...
D=6 N=4036 <=50K
Age
= Craft-repair, Transport-moving, ...
D=7 N=6293 <=50K
education-num
<= 40,500000
D=18 N=11604 <=50K
occupation
<= 33,500000
D=8 N=941 <=50K
> 33,500000
D=9 N=3095 >50K
education-num
<= 7,500000
D=12 N=1304 <=50K
> 7,500000
D=13 N=4989 <=50K
Age
= Prof-specialty, Protective-serv, ...
D=20 N=1568 <=50K
<= 9,500000
D=10 N=1432 <=50K
> 9,500000
D=11 N=1663 >50K
<= 35,500000
D=14 N=1691 <=50K
> 35,500000
D=15 N=3298 <=50K
Age

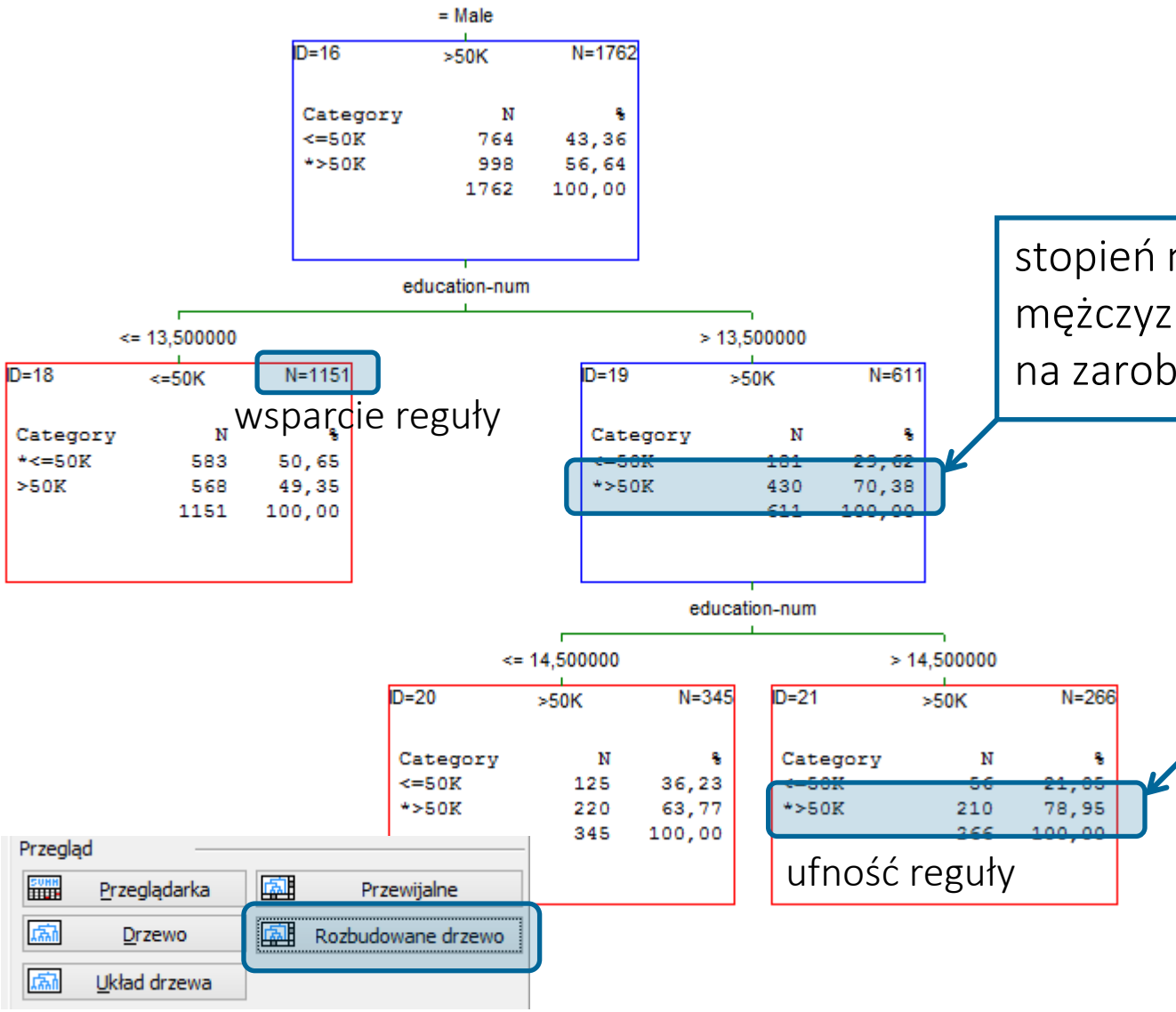

Wyniki drzew interakc. C&RT: adult
Menedżer
Podstawowe
Klasyfikacja
Predykcja
Raport
Drzewo (budowa, przycinanie):
Buduj drzewo
Eksploruj
Buduj i przytnij drzewo
Usuń wszystkie
Buduj 1 poziom
Usuń 1 poziom
Przegląd drzewa:
Przeglądarka
Przewijalne
Drzewo
Układ drzewa
Węzły i gałęzie:
Węzeł:
1
Określanie podziałów
Buduj gałąź
Buduj 1 poziom gałęzi
Stat. predyktorów
Usuń gałąź
Warunek podziału
Kod SQL
Dane
Histogram zm. zal.
Wybierz zastępcę
brak
Stat. zastępcy
Zapisz drzewo
Nowe
Zamknij
Opcje
Skoroszyt10* - węzeł 11 - ">50K"01
Skoroszyt10*
węzeł 1 - "<=50K"0
węzeł 2 - "<=50K"0
węzeł 4 - "<=50K"0
węzeł 6 - "<=50K"0
węzeł 8 - "<=50K"0
węzeł 9 - ">50K"01
węzeł 11 - ">50K"01
węzeł 12 - "<=50K"0
węzeł 13 - "<=50K"0
węzeł 14 - "<=50K"0
węzeł 5 - ">50K"01
węzeł 16 - "<=50K"0
węzeł 18 - "<=50K"0
węzeł 20 - "<=50K"0
węzeł 21 - "<=50K"0
węzeł 22 - "<=50K"0
węzeł 24 - "<=50K"0
węzeł 25 - "<=50K"0
węzeł 26 - "<=50K"0
węzeł 27 - "<=50K"0
węzeł 23 - "<=50K"0
węzeł 19 - "<=50K"0
węzeł 28 - "<=50K"0
węzeł 29 - "<=50K"0
węzeł 17 - "<=50K"0
węzeł 30 - "<=50K"0
węzeł 31 - "<=50K"0
węzeł 32 - "<=50K"0
węzeł 33 - "<=50K"0
Węzeł 11
Liczba przypadków w węźle: 1663
wybrana kategoria: ">50K"
666
997
1200
1000
800
600
400
200
0
<=50K
>50K
węzeł 10 - "<=50K"0
węzeł 11 - ">50K"01

Reguły

- Jeżeli osoba pozostaje w związku małżeńskim i jej liczba lat edukacji przekracza 12,5 roku, wtedy jej dochód prawdopodobnie przekracza 50 000 $ (węzeł ID5)
  (z prawdopodobieństwem… 72%)
- Jeżeli osoba pozostaje w związku małżeńskim, jej liczba lat edukacji nie przekracza 12,5 roku, wykonuje zawód… oraz ma ponad 33,5 lat wtedy jej dochód prawdopodobnie przekracza 50 000 $ (węzeł ID9)
  (z prawdopodobieństwem… 53%)
- Jeżeli osoba ma ponad 33,5 lat, pozostaje w związku małżeńskim, liczba lat jej edukacji mieści się w przedziale 9,5 do 12,5 lat, wykonuje zawód… wtedy jej dochód prawdopodobnie przekracza 50 000 $ (węzeł ID11)
  (z prawdopodobieństwem… 60%)



Pewność reguł (ufność reguły, dokładność)



stopień magistra daje mężczyznom 70,38% szans na zarobki pow.50tys

studia podyplomowe i doktorat podnosi szansę na zarobki pow.50tys o ponad 8 punktów procentowych

Wsparcie i Ufność

jest bardzo mało kobiet wysoko wykształconych

ID=1 <=50K N=9950

Category	N	%
*<=50K	7540	75,78
>50K	2410	24,22
	9950	100,00

Wsparcie = 217/9950 = 2,2%

= Female

ID=17 <=50K N=697

Category	N	%
*<=50K	518	74,32
>50K	179	25,68
	697	100,00

większość kobiet, nawet bardzo dobrze wykształconych, nie zarabia pow. 50 tys.

education-num

<= 13,500000

ID=22 <=50K N=480

Category	N	%
*<=50K	390	81,25
>50K	90	18,75
	480	100,00

> 13,500000

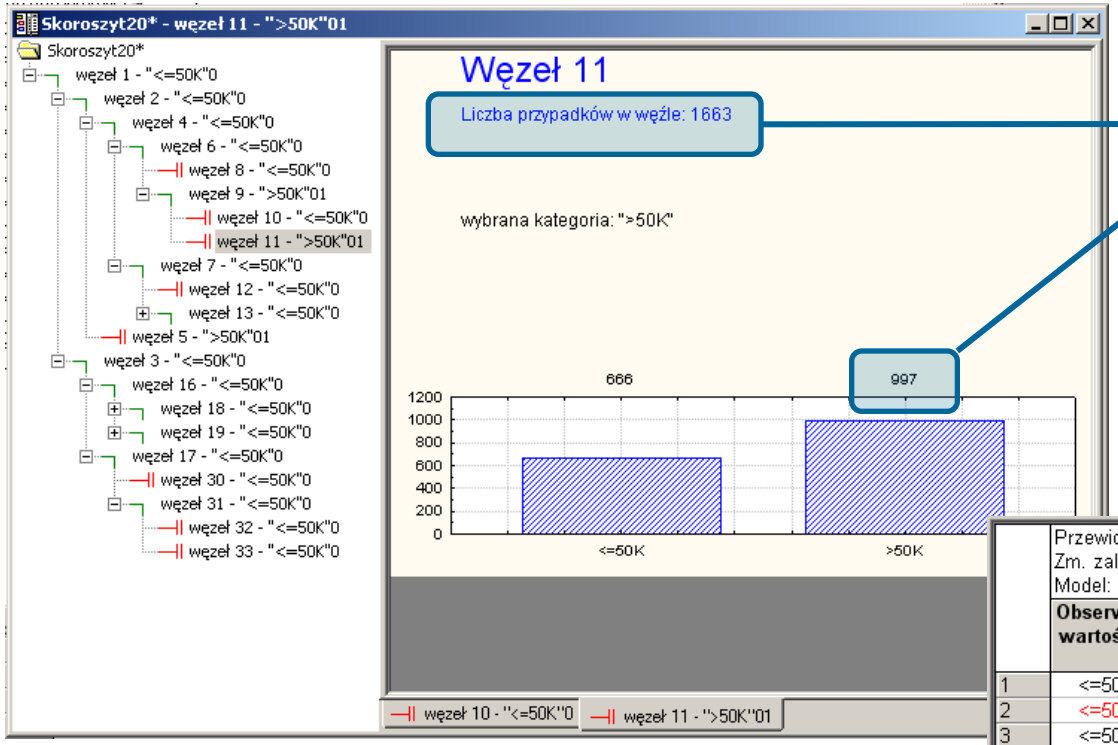
ID=23 <=50K N=217

Category	N	%
*<=50K	128	58,99
>50K	89	41,01
	217	100,00

Ufność = 41,01%

kobiety ze stopniem magistra (i wyżej) mają jedynie 41% szans na zarobki >50K

Wsparcie (pokrycie) reguły / ufność (dokładność)



ID11: 997/1633 = 0,5995

ufność reguły

$N_{konkluzji} / N_{węzła}$

Przewidywane (adult)
Zm. zal.:Income
Model: C&RT

	Obserw. wartość	Przewid. wartość	Prawdopodobieństwo <=50K	Prawdopodobieństwo >50K	Końcowe węzły
1	<=50K	<=50K	0,814680	0,185320	32
2	<=50K	>50K	0,276399	0,723601	5
3	<=50K	<=50K	0,993289	0,006711	27
4	<=50K	<=50K	0,900307	0,099693	12
5	<=50K	>50K	0,276399	0,723601	5
6	<=50K	>50K	0,276399	0,723601	5
7	<=50K	<=50K	0,967155	0,032845	23
8	>50K	<=50K	0,560754	0,439246	10
9	>50K	<=50K	0,577596	0,422404	33
10	>50K	>50K	0,276399	0,723601	5
11	>50K	>50K	0,400481	0,599519	11
12	>50K	>50K	0,276399	0,723601	5
13	<=50K	<=50K	0,939372	0,060628	30
14	<=50K	<=50K	0,942011	0,057989	28
15	>50K	<=50K	0,660703	0,339297	15
16	<=50K	<=50K	0,900307	0,099693	12
17	<=50K	<=50K	0,997584	0,002416	24
18	<=50K	<=50K	0,993289	0,006711	27

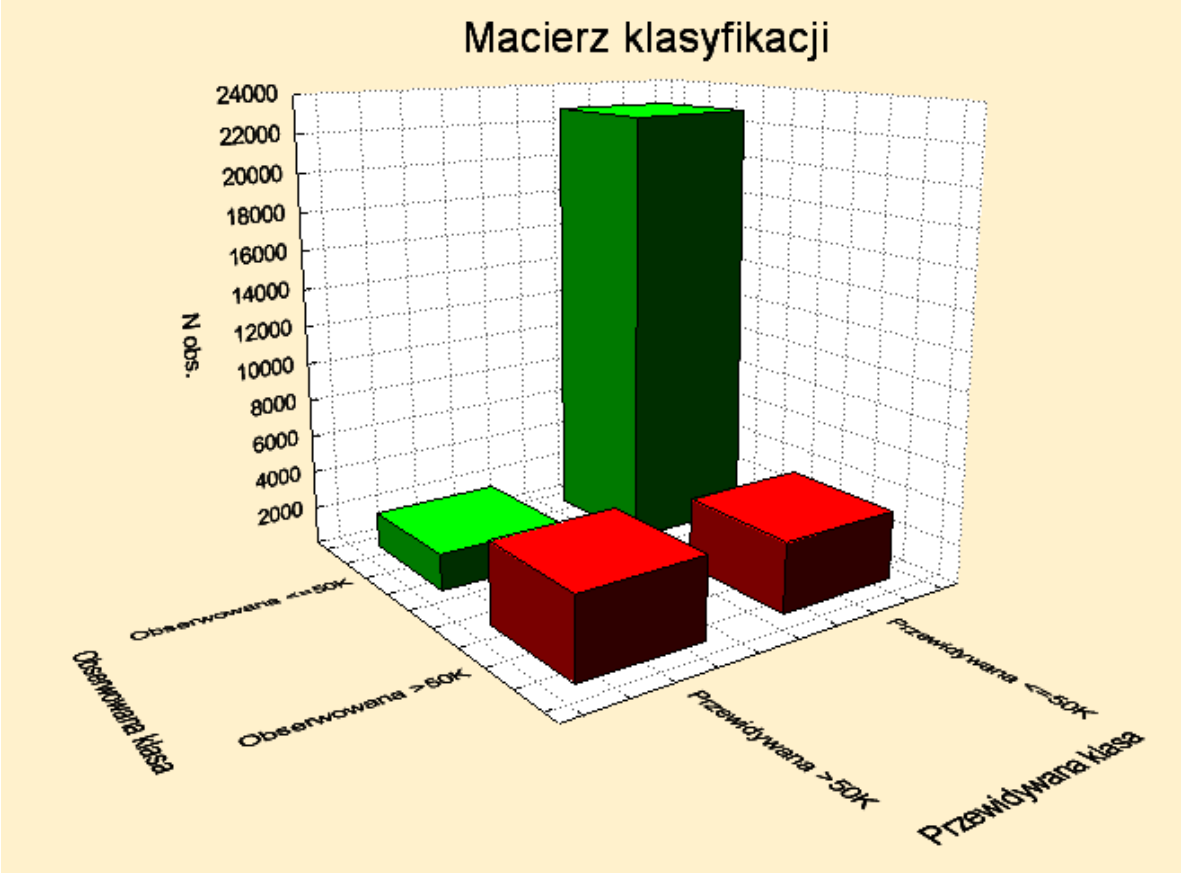
Macierz klasyfikacji (adult) | Macierz klasyfikacji (adult) | Przewidywane (adult) | Wykres przyros

wsparcie reguły:

$N_{węzła} / N_{zbioru}$

ID11: 1633/32561=5%

Macierz klasyfikacji



Ile razy model
się pomylił?
Pojęcie „kosztu”

Macierz klasyfikacji (adult)
Zm. zal.:Income
Model: C&RT

	Obserw.	Przewidywana <=50K	Przewidywana >50K	Łącznie w wierszu
Liczba	<=50K	22829	1891	24720
Procent z kolumnny		86.26%	31.03%	
Procent z wiersza		92.35%	7.65%	
Procent z ogółu		70.11%	5.81%	75.92%
Liczba	>50K	3637	4204	7841
Procent z kolumnny		13.74%	68.97%	
Procent z wiersza		46.38%	53.62%	
Procent z ogółu		11.17%	12.91%	24.08%
Liczba	Ogół grup	26466	6095	32561
Procent łącznie		81.28%	18.72%	

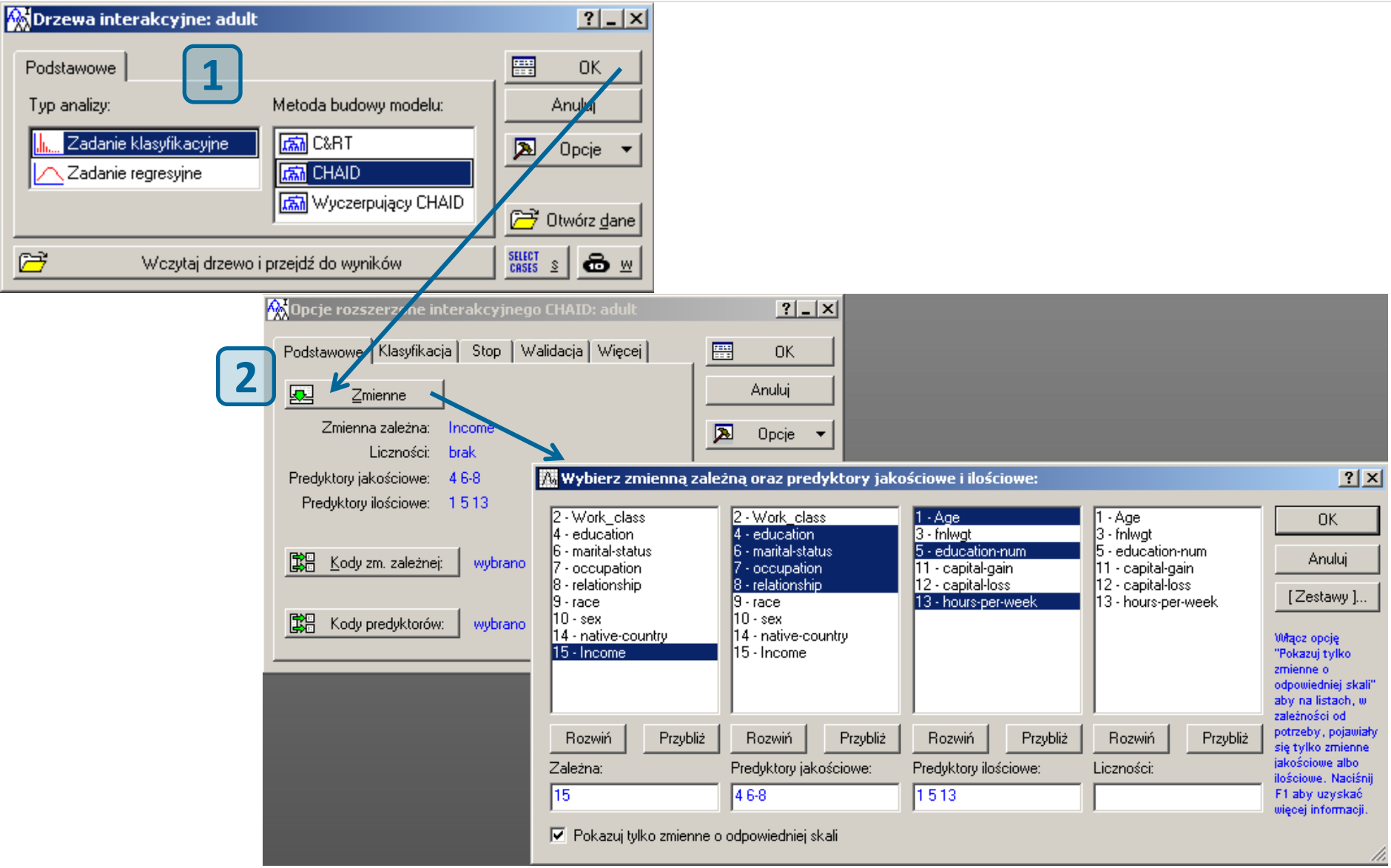
FP

FN

CHAID

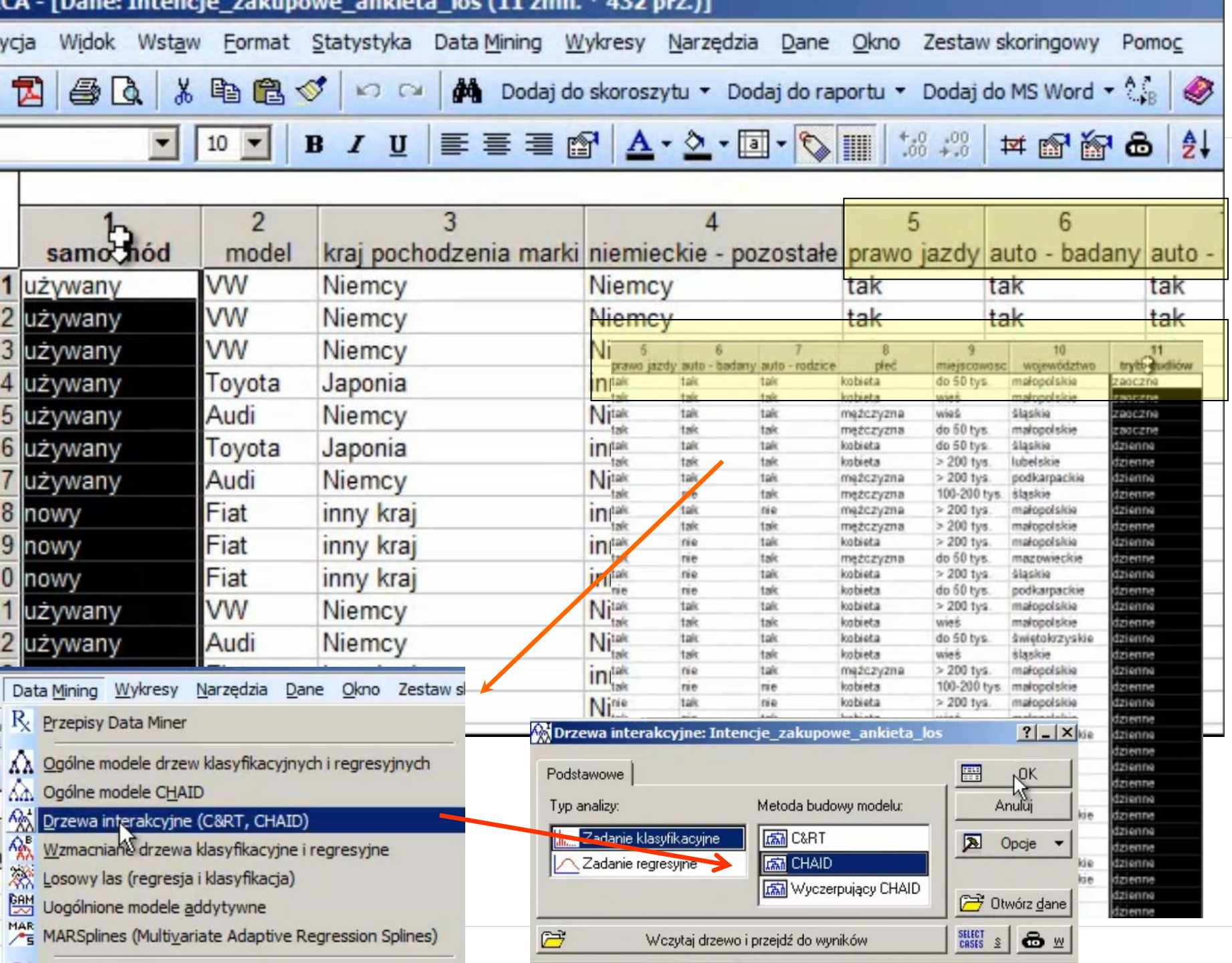
- Algorytm drzew klasyfikacyjnych
- Zmienne ilościowe dzielone są na 10 kategorii, zmienne jakościowe obsługiwane w sposób naturalny
- Wyszukiwanie par kategorii podobnych do siebie ze względu na zmienną zależną
- Test Chi^2

STATISTICA - przykład drzewa CHAID



Przykład: badanie intencji zakupowych studentów

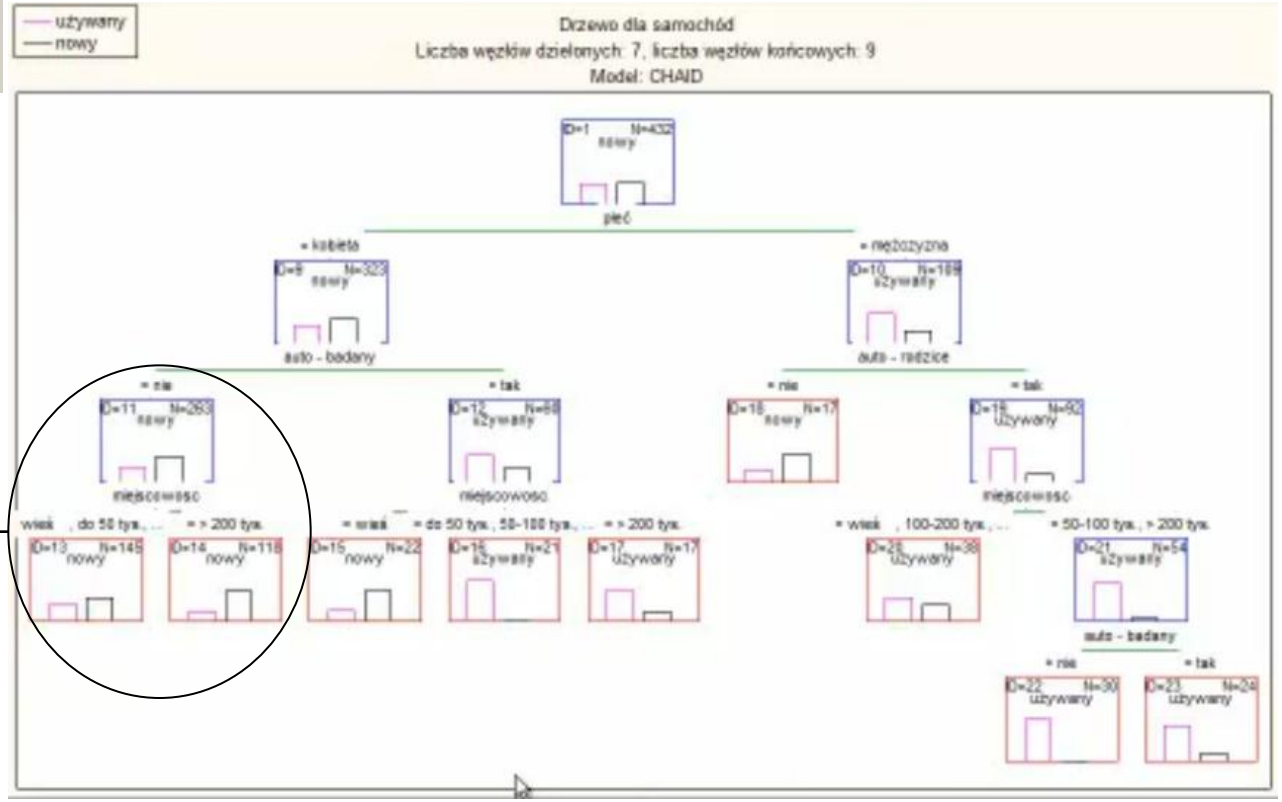
- Co wpływa na skłonność zakupu samochodu nowego bądź używanego?
- Wybór jednego z 12 profili aut o porównywalnej cenie (połowa z nich używane, połowa – nowe)
- 1200 ankietowanych,
- dane demograficzne + wybór

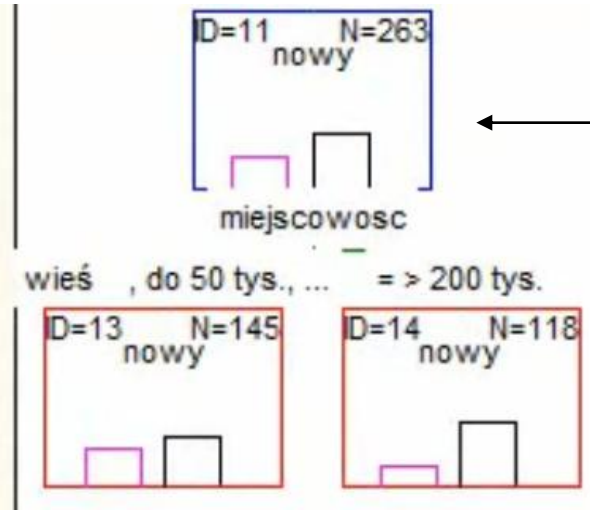

CA - [Dane: Intencje_zakupowe_ankieta_los (11 zmi. * 432 prz.)]
ycja Widok Wstaw Format Statystyka Data Mining Wykresy Narzędzia Dane Okno Zestaw skoringowy Pomoc
Dodaj do skoroszytu
Dodaj do raportu
Dodaj do MS Word
Data Mining Wykresy Narzędzia Dane Okno Zestaw s
Przepisy Data Miner
Ogólne modele drzew klasyfikacyjnych i regresyjnych
Ogólne modele CHAID
Drzewa interakcyjne (C&RT, CHAID)
Wzmacniane drzewa klasyfikacyjne i regresyjne
Losowy las (regresja i klasyfikacja)
Uogólnione modele addytywne
MARSplines (Multivariate Adaptive Regression Splines)
Drzewa interakcyjne: Intencje_zakupowe_ankieta_los
Podstawowe
Typ analizy:
Zadanie klasyfikacyjne
Zadanie regresyjne
Metoda budowy modelu:
C&RT
CHAID
Wyczerpujący CHAID
OK
Anuluj
Opcje
Otwórz dane
Wczytaj drzewo i przejdź do wyników

Model: CHAID

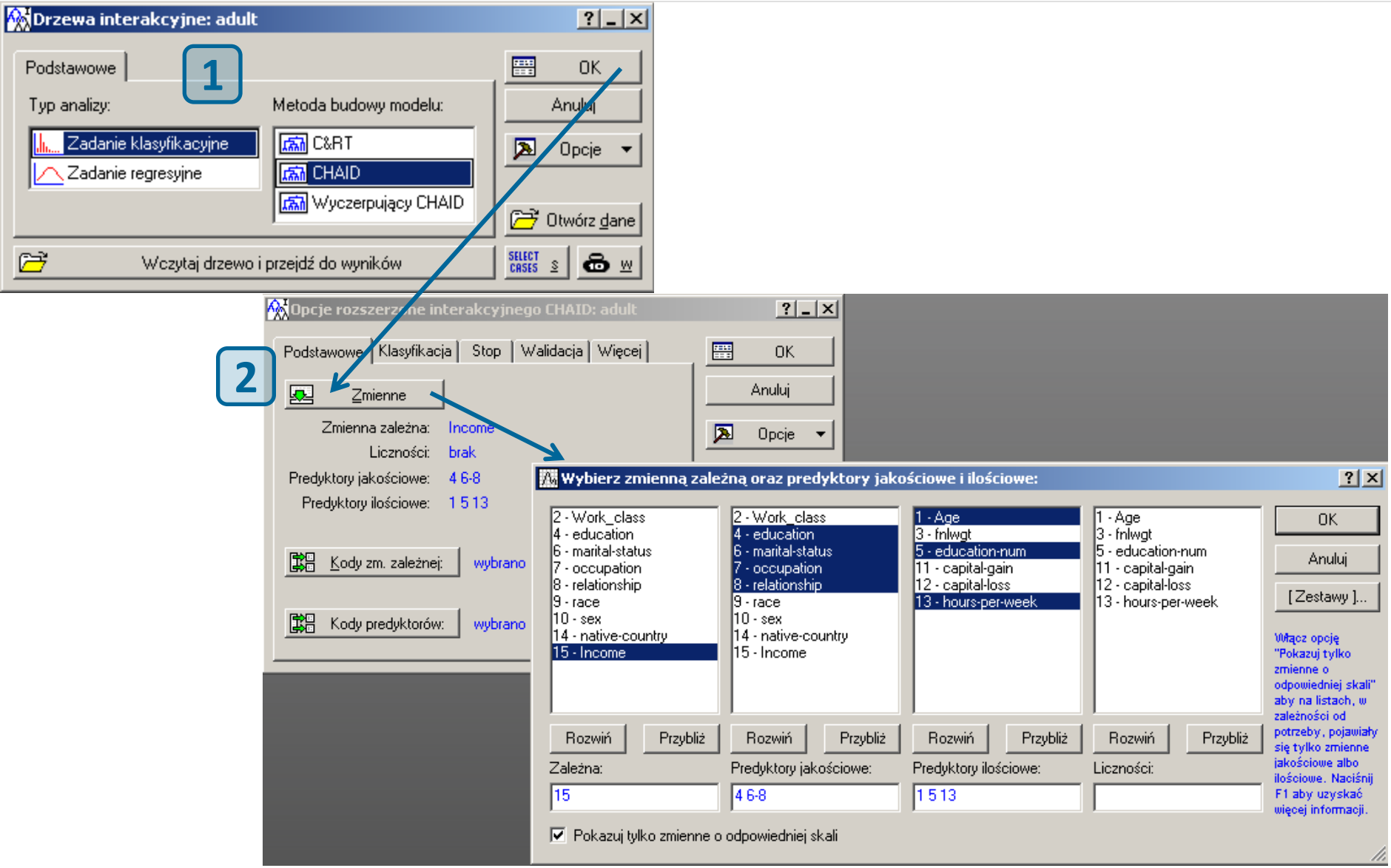
	Liczba węzły	Typ podziału	chi-kwadrat Statystyka	Stopnie Swobody	Skorygowane p
płeć	2	Automatycznie	27,13573	1,000000	0,000000
prawo jazdy	2	Automatycznie	21,86719	1,000000	0,000003
auto - badany	2	Automatycznie	17,78424	1,000000	0,000025
tryb studiów	3	Automatycznie	28,19740	2,000000	0,000001
województwo	2	Automatycznie	12,80863	1,000000	0,022082
miejscowosc	2	Automatycznie	11,17302	1,000000	0,013279
auto - rodzice	2	Automatycznie	3,89359	1,000000	0,048471

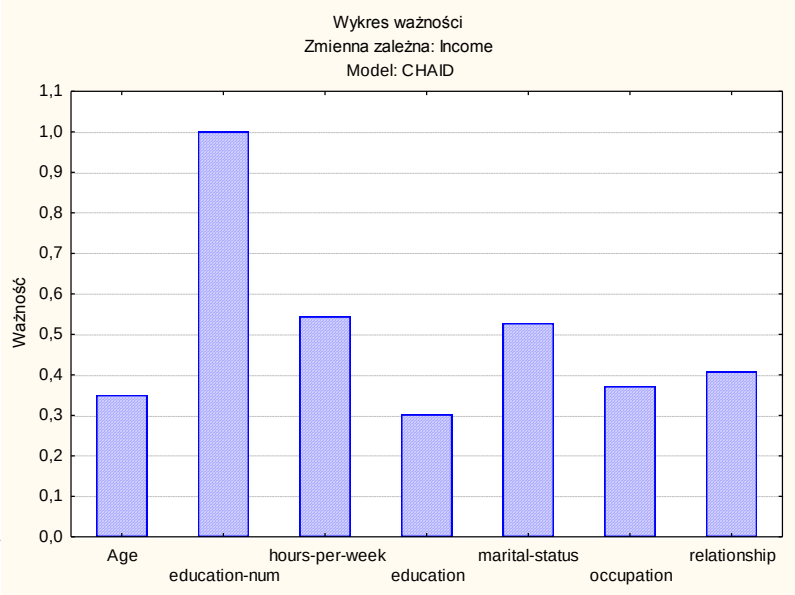
Zmienną decyzyjną najsilniej różnicuje płeć



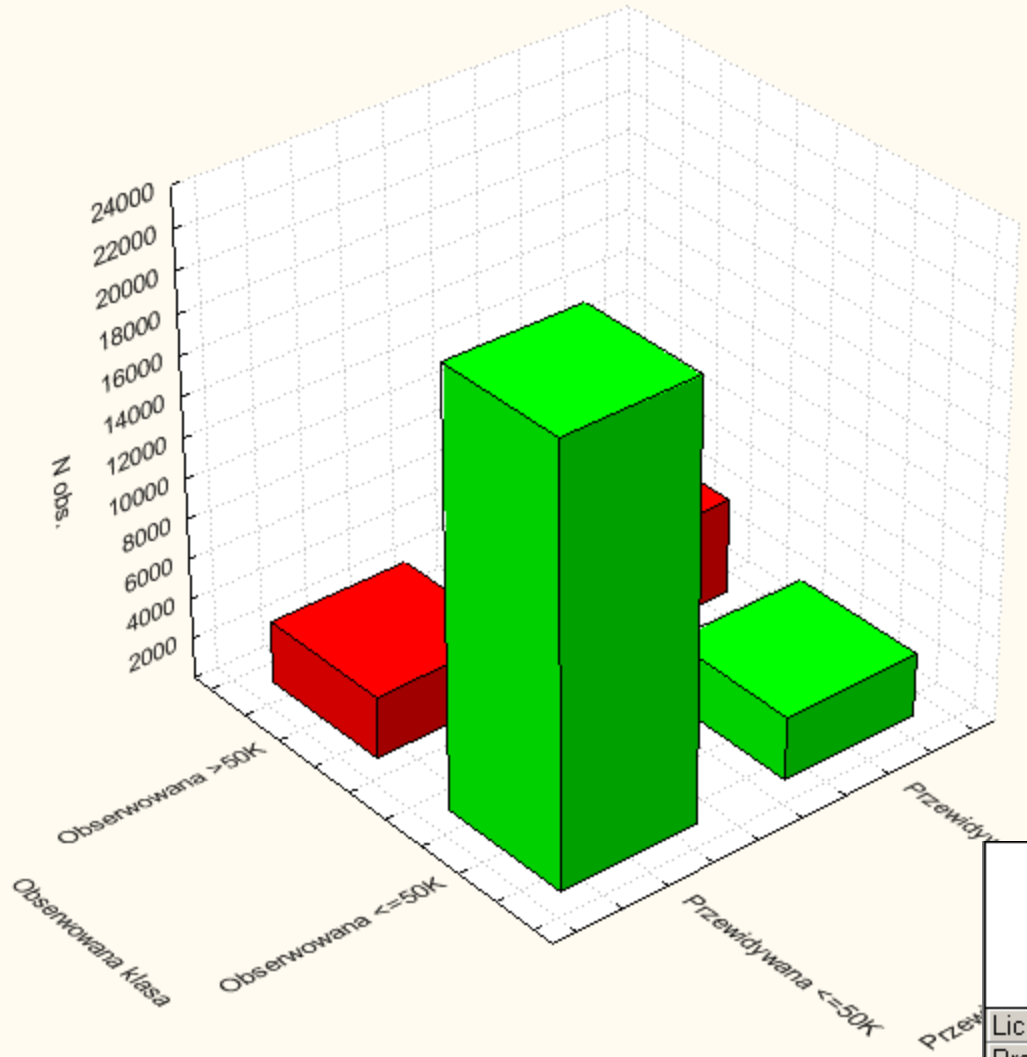


STATISTICA - inny przykład drzewa CHAID




Wykres ważności
Zmienna zależna: Income
Model: CHAID
Ważność
1,1
1,0
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0,0
Age
education-num
hours-per-week
education
marital-status
occupation
relationship

Macierz klasyfikacji



	Macierz klasyfikacji (adult) Zm. zal.:Income Model: CHAID			
	Obserw.	Przewidywana <=50K	**Przewidywana >50K**	Łącznie w wierszu
Liczba	<=50K	21569	3151	24720
Procent z kolumnny		87.63%	39.65%	
Procent z wiersza		87.25%	12.75%	
Procent z ogółu		66.24%	9.68%	75.92%
Liczba	>50K	3045	4796	7841
Procent z kolumnny		12.37%	60.35%	
Procent z wiersza		38.83%	61.17%	
Procent z ogółu		9.35%	14.73%	24.08%
Liczba	Ogół grup	24614	7947	32561
Procent łącznie		75.59%	24.41%	

FP 2

FN 1

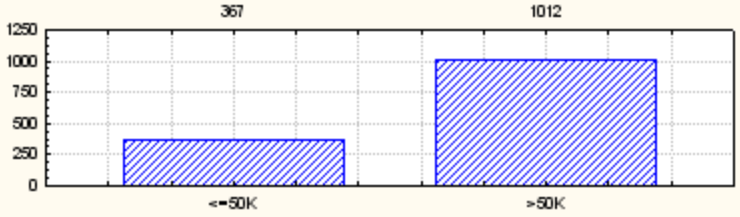
Skoroszyt22* - węzeł 17 - ">50K"01

Skoroszyt22*
- węzeł 1 - "<=50K"0
  - węzeł 2 - "<=50K"0
    - węzeł 6 - "<=50K"0
      - węzeł 8 - "<=50K"0
      - węzeł 9 - "<=50K"0
        - węzeł 10 - "<=50K"0
        - węzeł 11 - "<=50K"0
    - węzeł 7 - "<=50K"0
  - węzeł 3 - "<=50K"0
    - węzeł 12 - ">50K"01
      - węzeł 16 - "<=50K"0
        - węzeł 18 - "<=50K"0
          - węzeł 20 - "<=50K"0
          - węzeł 21 - ">50K"01
        - węzeł 19 - ">50K"01
      - węzeł 17 - ">50K"01
    - węzeł 13 - "<=50K"0
      - węzeł 22 - "<=50K"0
      - węzeł 23 - "<=50K"0
      - węzeł 24 - "<=50K"0
    - węzeł 14 - ">50K"01
    - węzeł 15 - "<=50K"0
      - węzeł 25 - "<=50K"0
      - węzeł 26 - ">50K"01
  - węzeł 4 - "<=50K"0
    - węzeł 27 - "<=50K"0
    - węzeł 28 - "<=50K"0
    - węzeł 29 - "<=50K"0
  - węzeł 5 - "<=50K"0

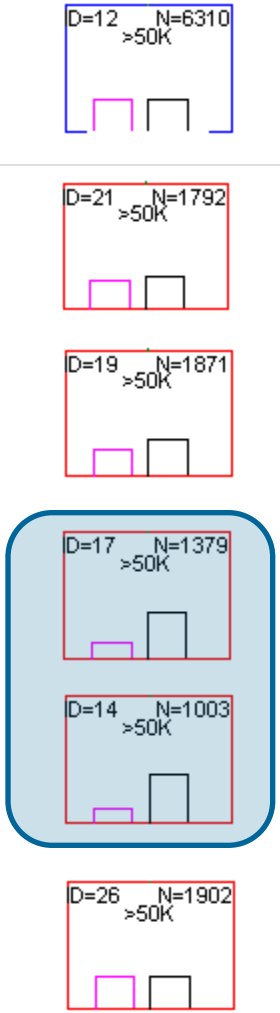
Węzeł 17

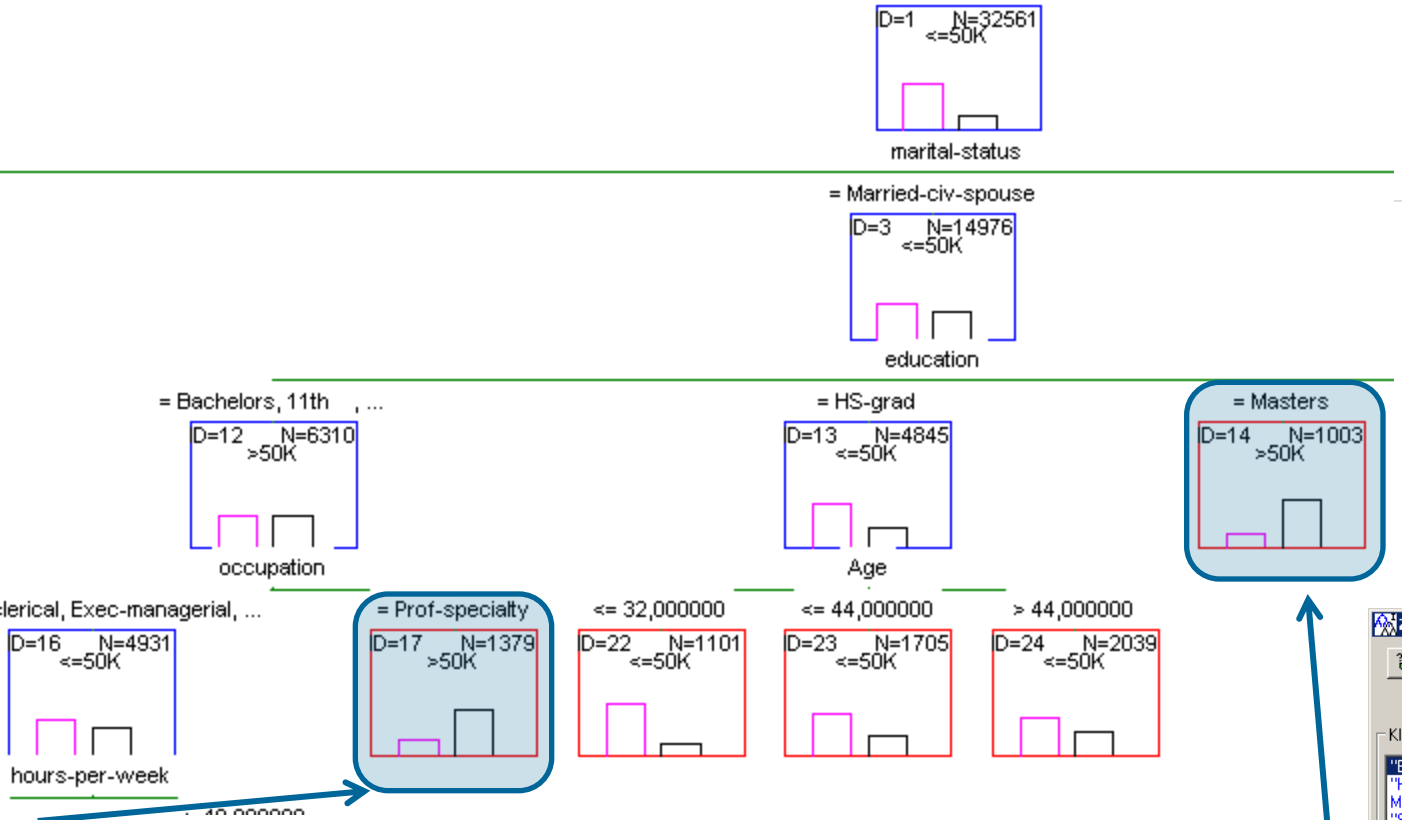
Liczba przypadków w węźle: 1379

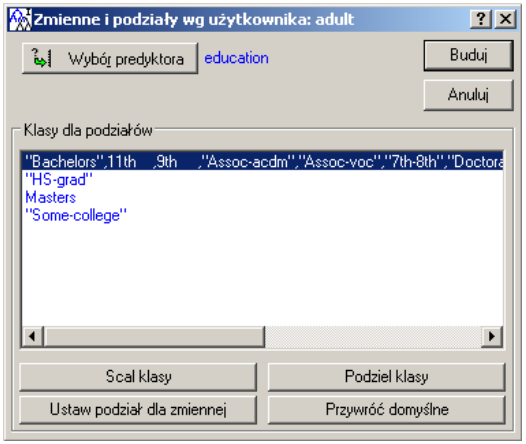
wybrana kategoria: ">50K"



węzeł 16 - "<=50K"0 | węzeł 17 - ">50K"01







Jeżeli osoba pozostaje w związku małżeńskim skończyła szkołę z grupy…, ale jest profesjonalistą w swoim zawodzie, wtedy jej dochód prawdopodobnie przekracza 50 000 $ (węzeł ID17) (z prawdopodobieństwem… 73%)

Jeżeli osoba pozostaje w związku małżeńskim i skończyła studia magisterskie, wtedy jej dochód prawdopodobnie przekracza 50 000 $ (węzeł ID14) (z prawdopodobieństwem… 77%)

Drzewa regresyjne

Dla drzew klasyfikacyjnych stosuje się różne miary niejednorodności: Indeks Giniego, Chi-kwadrat lub G-kwadrat.

Podział węzłów w drzewach regresyjnych, następuje na podstawie odchylenia najmniejszych kwadratów (LSD - *Least Significant Difference*).

$$R(t) = \frac{1}{N_w(t)} \sum_{i \in t} w_i f_i (y_i - \bar{y}(t))^2$$

gdzie

$N_w(t)$ - ważona liczba przypadków w węźle t,

w_i - wartość zmiennej ważącej dla przypadku i,

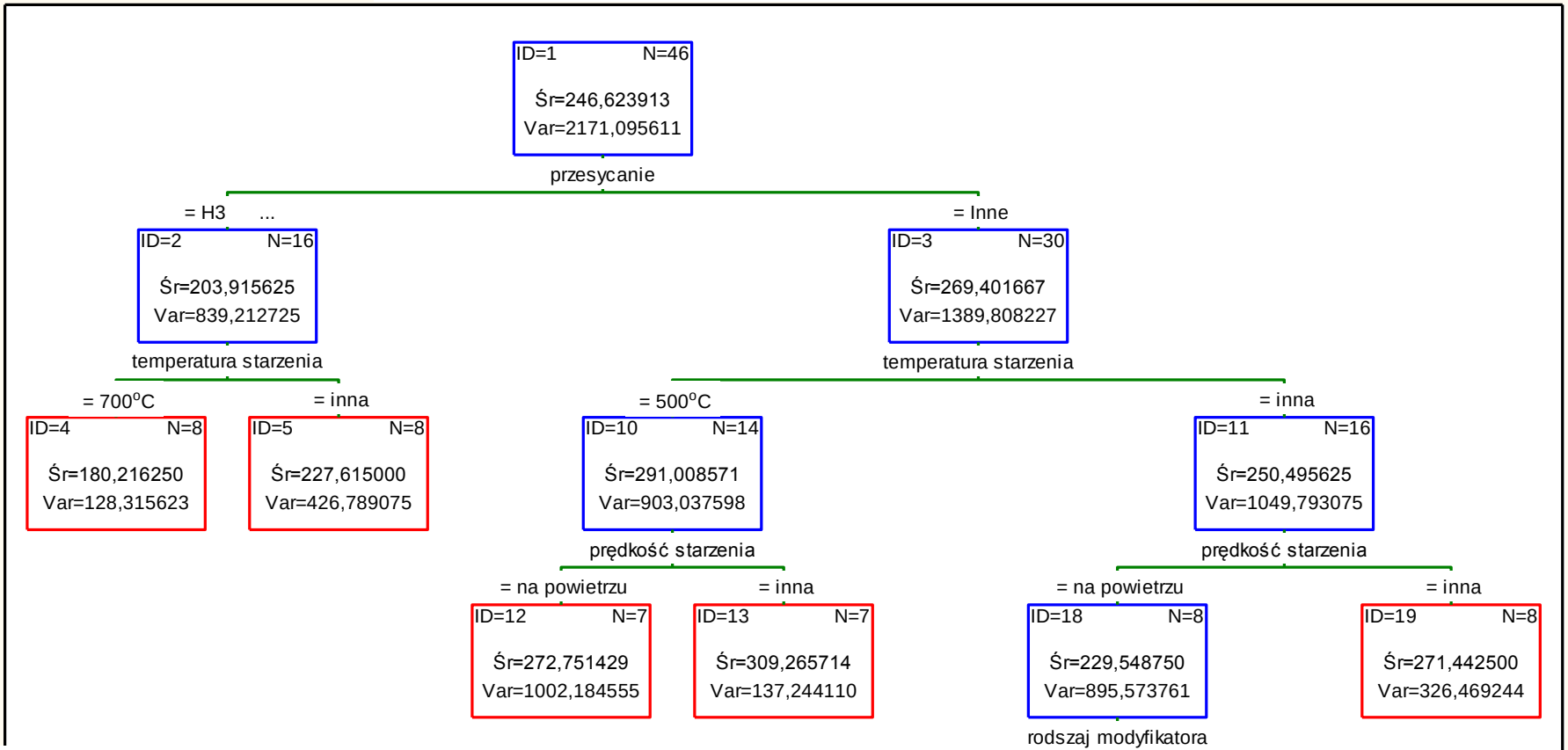
f_i - wartość zmiennej częstotliwości,

y_i - wartość zmiennej odpowiedzi,

y(t) jest średnią ważoną w węźle t.

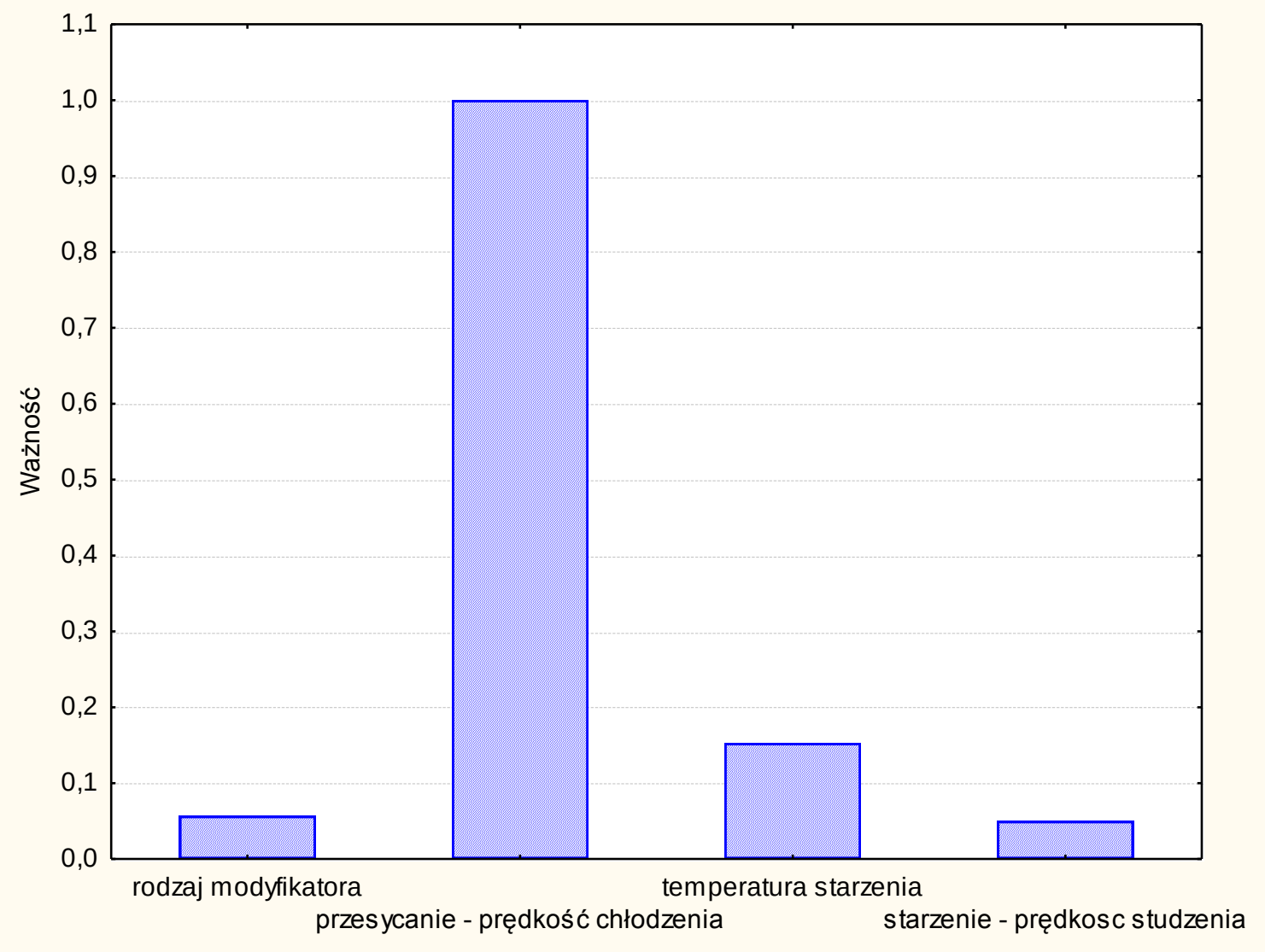
Źródło: dla wzorów wykorzystywanych przez model *C&RT* zaimplementowany w *STATISTICA* wykorzystano fragmenty z Internetowego Podręcznika Statystyki, StatSoft, Inc., 1984-2005, jest to oficjalny podręcznik wydany przez dystrybutora oprogramowania.

Drzewo dla parametru: umowna granica plastyczności $R_{0,2}$



Klasy dla poszczególnych parametrów R_m, $R_{0,2}$, A zostały wyznaczone za pomocą modeli drzew regresyjnych w oparciu o zmienne predykcyjne jakimi były:

- Rodzaj modyfikatora
- Przesycanie – prędkość chłodzenia
- Temperatura starzenia
- Starzenie – prędkość studzenia


1,1
1,0
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0,0
Ważność
rodzaj modyfikatora
przesycanie - prędkość chłodzenia
temperatura starzenia
starzenie - prędkosc studzenia

Ocena drzewa

- Dla potrzeb oceny modeli wprowadzono pojecie kosztu.
- Koszt określony jest poprzez wariancję.
- Konieczność minimalizacji kosztów wynika z tego, że niektóre błędy mogą mieć bardziej katastrofalne skutki niż inne.
- Jakość modelu regresyjnego oceniamy również poprzez współczynnik determinacji.

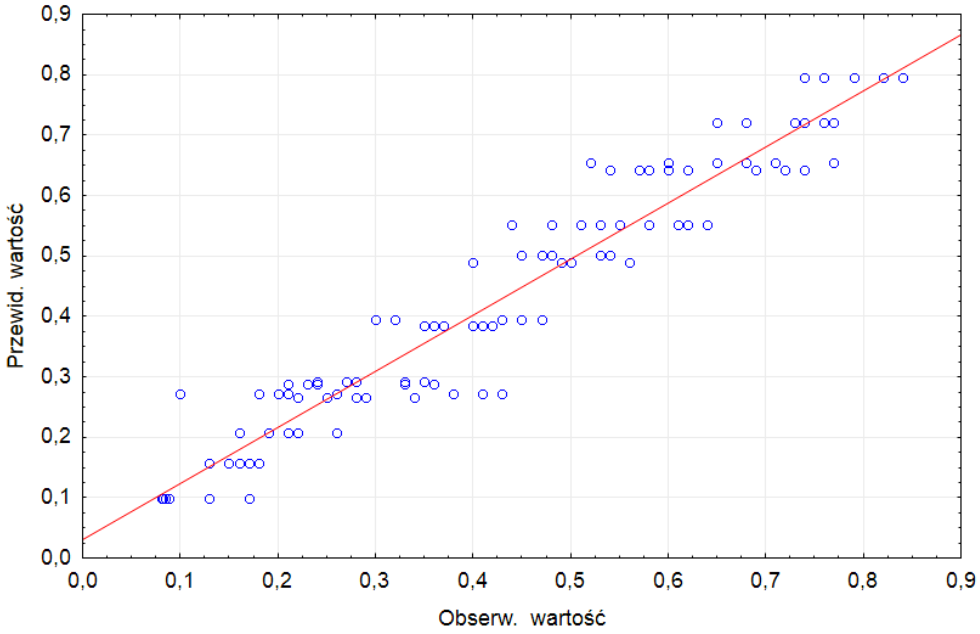
Współczynnik determinacji

r^2 (R^2) – współczynnik determinacji
(wielkość ta oznacza kwadrat współczynnika korelacji)
przyjmuje wartości z przedziału $[0,1]$
jest miarą stopnia w jakim model wyjaśnia kształtowanie się zmiennej Y.

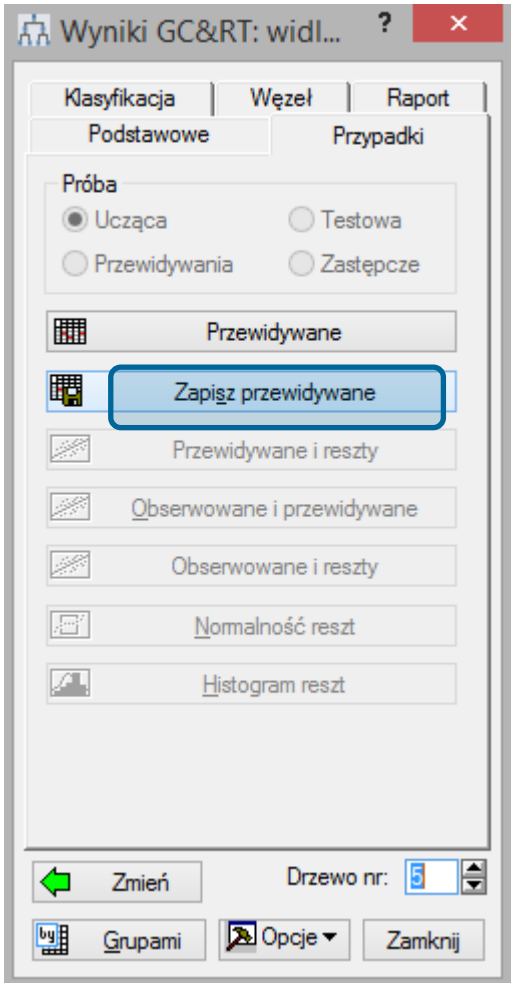
$$R^2 = \frac{\sum_{t=1}^{n}(\hat{y}_t - \overline{y})^2}{\sum_{t=1}^{n}(y_t - \overline{y})^2},$$

Jeśli wartość R^2 jest duża, to oznacza to, że błędy dla tego modelu są stosunkowo małe i w związku z tym model jest dobrze dopasowany do rzeczywistych danych.

Im jego wartość jest bliższa 1, tym lepsze dopasowanie modelu do danych empirycznych

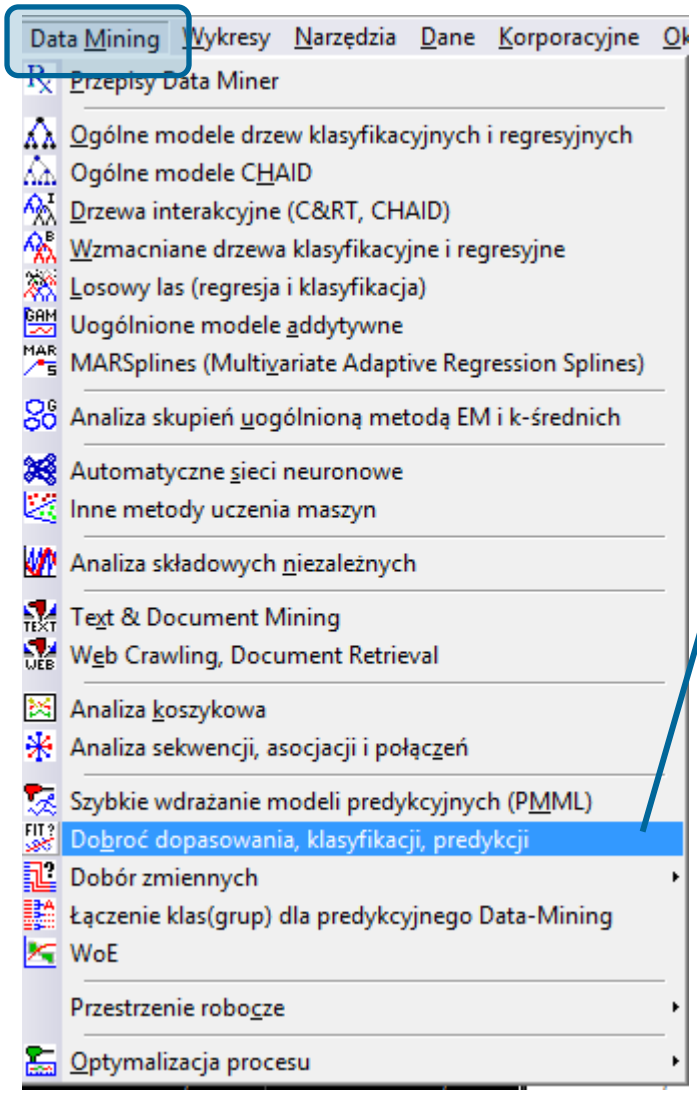


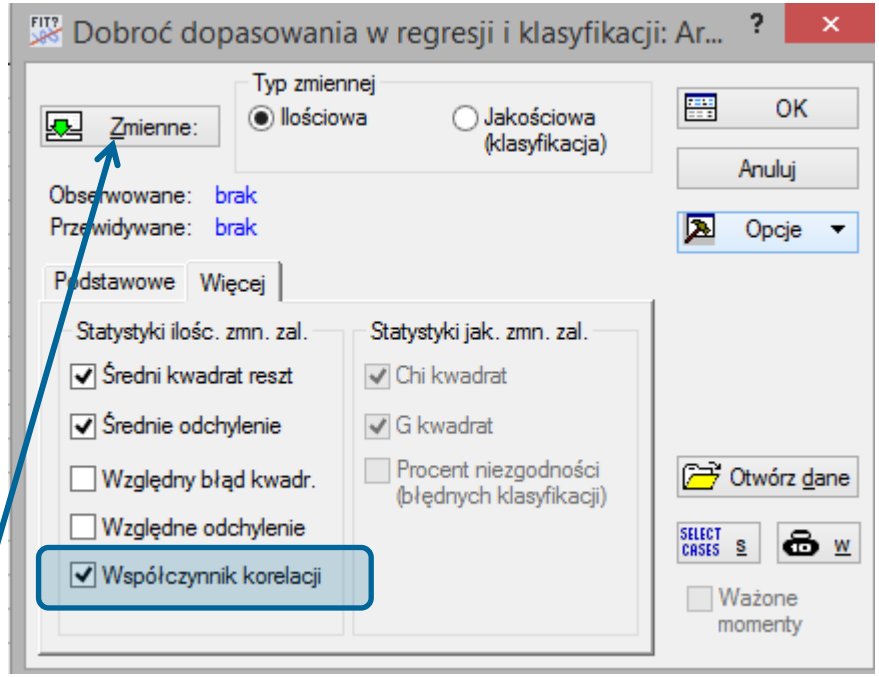
Dobroć dopasowania (1)



5 Obserw. wartość	6 Przewid. wartość
0,35000	0,27875
0,71000	0,73944
0,72000	0,73944
0,73000	0,73944
0,74000	0,73944
0,74500	0,73944
0,75000	0,73944
0,78000	0,82000
0,80000	0,82000
0,82000	0,82000
0,83000	0,82000
0,84000	0,82000
0,85000	0,82000
0,70000	0,73944
0,80000	0,73944
0,90000	0,96889
1,06000	0,96889

Dobroć dopasowania (2)



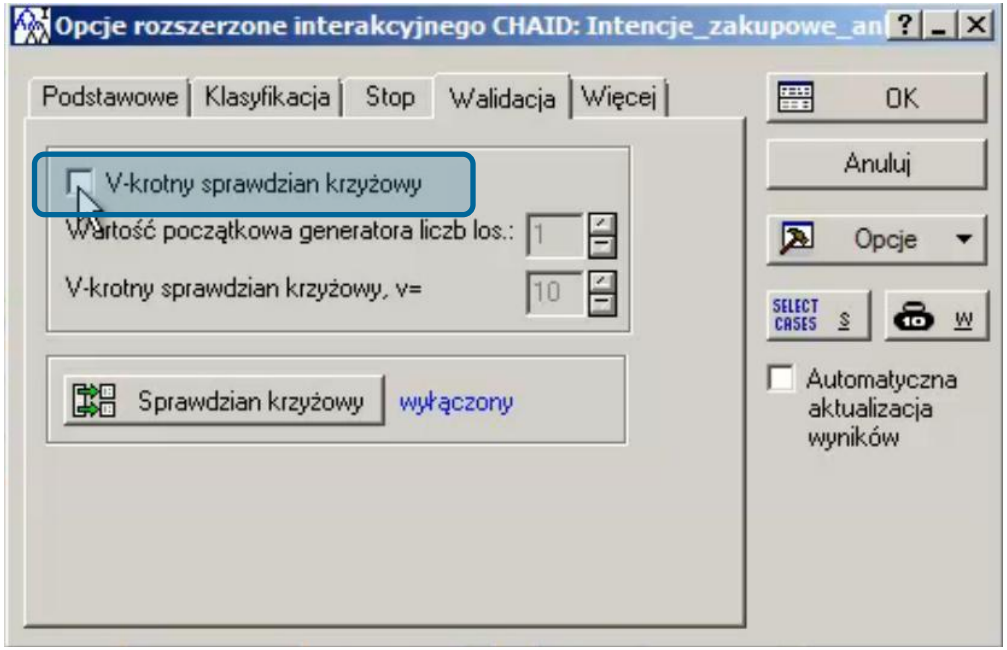


współczynnik determinacji
to kwadrat współczynnika korelacji

Koszt

Korzystamy z dwóch rodzajów kosztów: kosztu sprawdzianu krzyżowego (SK) oraz kosztu resubstytucji.

Wybiera się drzewo o minimalnym koszcie SK, lub drzewo najmniej złożone, którego koszty SK nie różnią się „znacznie” od minimalnych



Koszt resubstytucji

Narzędziem pomocniczym jest **koszt resubstytucji**.

Oblicza się tu oczekiwany błąd kwadratowy dla próby uczącej.

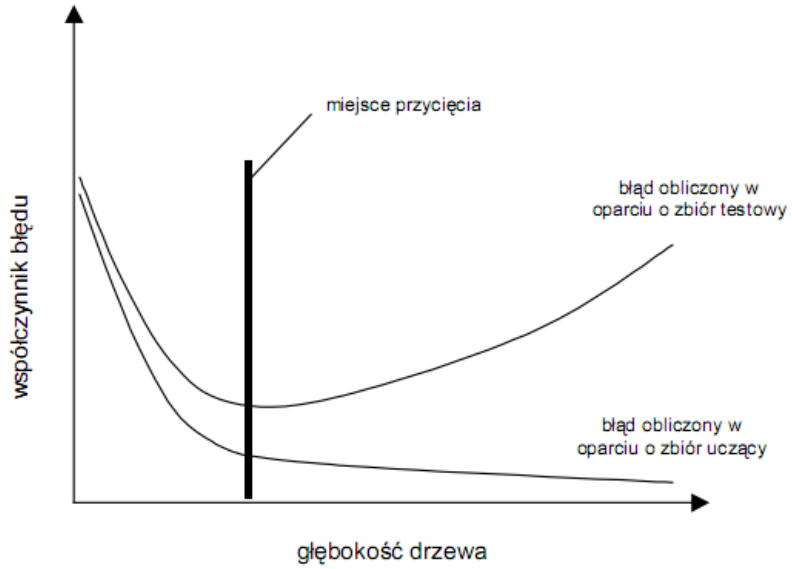
$$R(d) = \frac{1}{N}\sum_{i=1}^{N}(y_i - d(x_i))^2$$

gdzie próba ucząca *Z* składa się z punktów (x_i,y_i), i = 1,2,...,N.

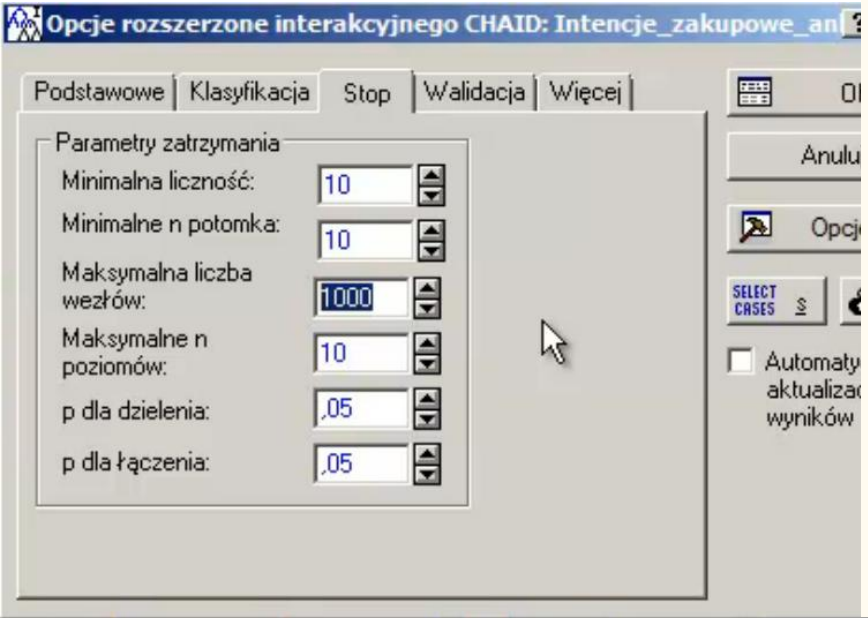
Miara ta obliczana jest dla tego samego zbioru danych, na bazie którego zbudowano model (partycję) d.

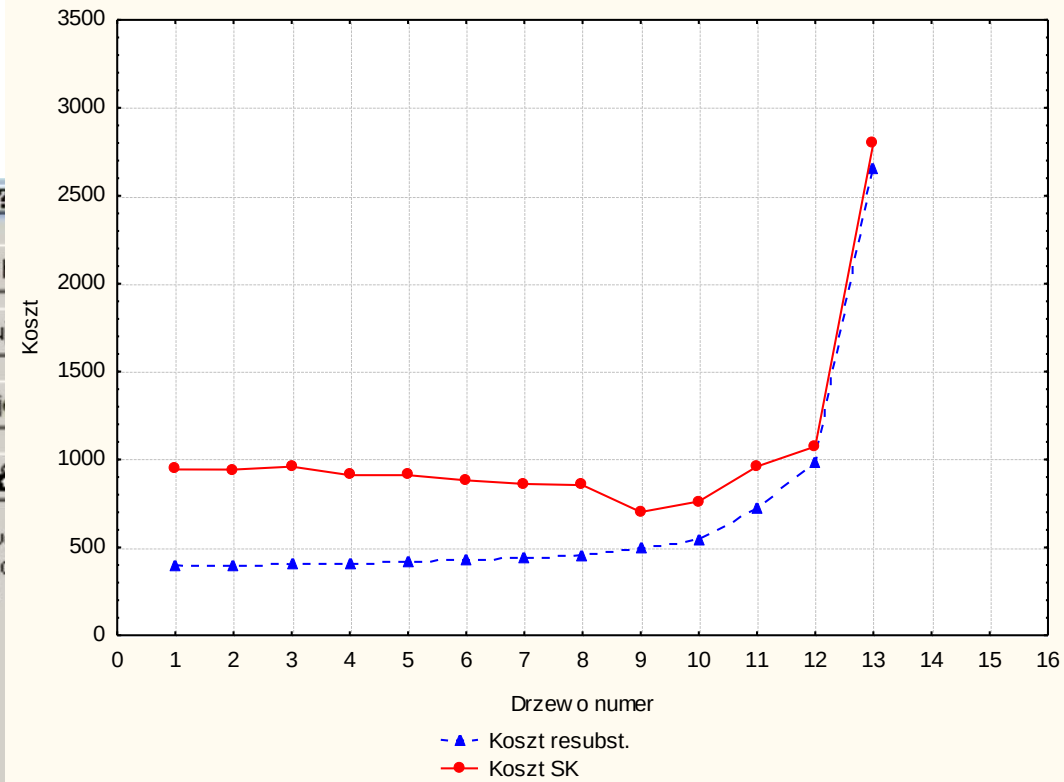
niski koszt resubstytucji = wartości zmiennej zależnej bliskie średniej w danym liściu

Wybór drzewa – przycinanie drzewa (1)

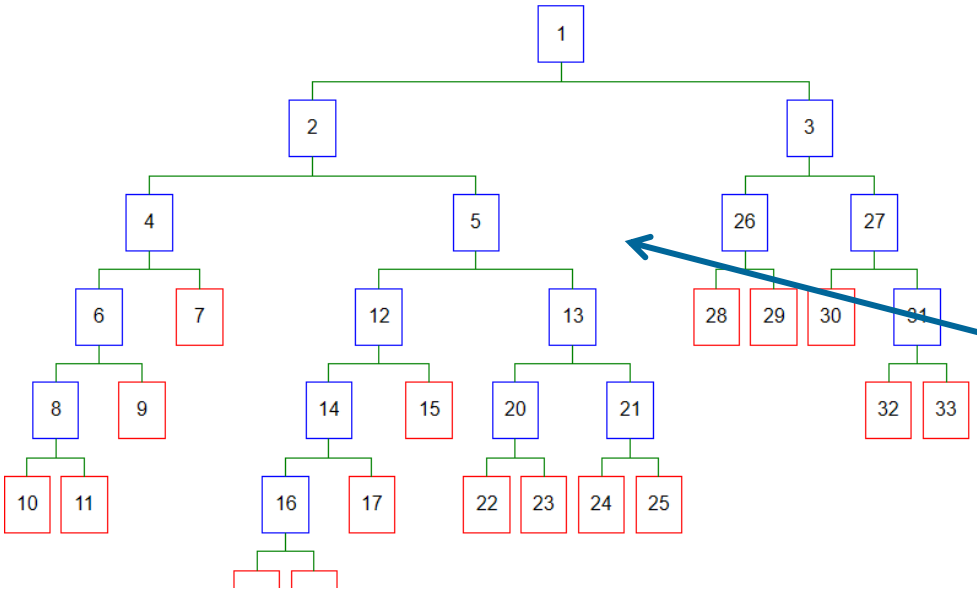


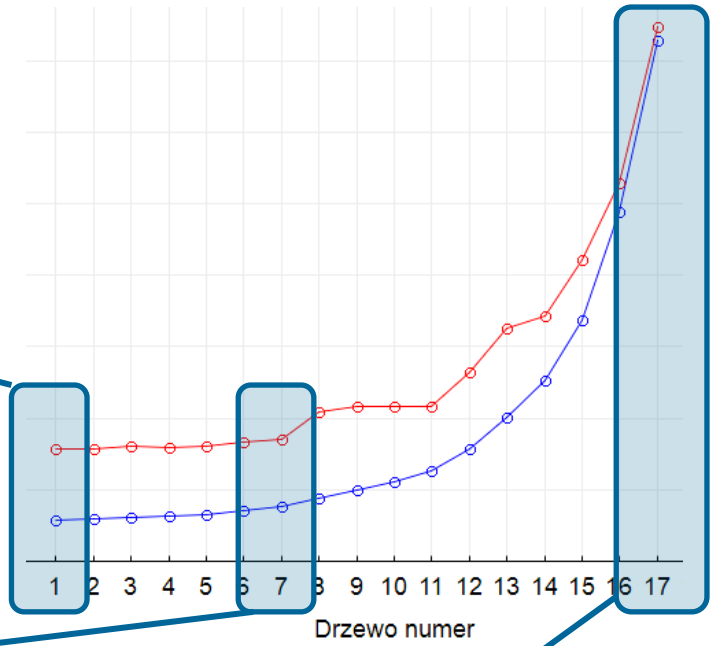
Jedną z metod doboru drzewa jest wybranie takiego, dla którego koszty resubstytucji i koszt sprawdzianu krzyżowego (SK) się przecinają.

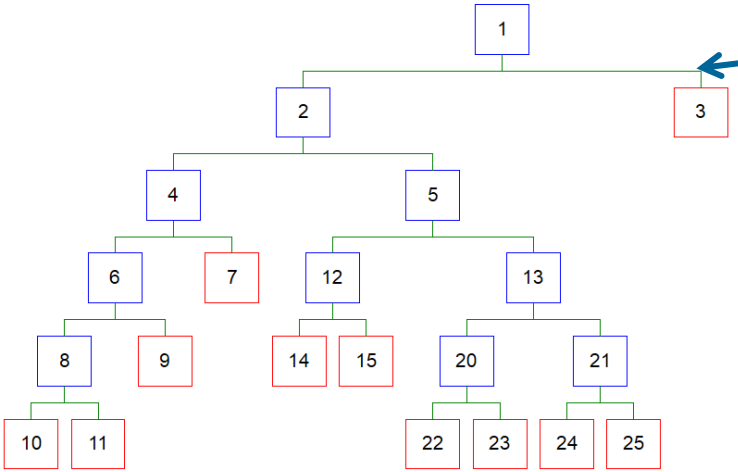


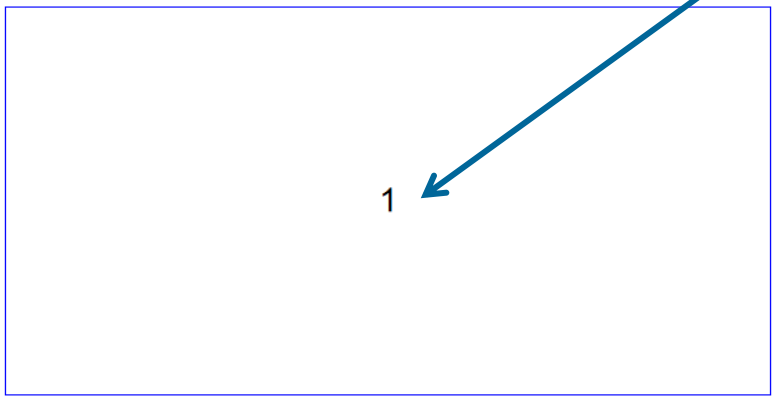


Wybór drzewa – przycinanie drzewa (2)









Efekt?

na podstawie drzewa nr 9 dla R_m można określić reguły:

- Jeśli próbka poddana została przesycaniu H3 i starzeniu w 500°C, wtedy wytrzymałość będzie miała rozkład o średniej *E(X)*=476[Mpa] i wariancji $D^2(X)$=793
- Jeśli próbka poddana została przesycaniu H3 i starzeniu w 700°C lub bez starzenia, wtedy wytrzymałość będzie miała rozkład o średniej *E(X)*=530[Mpa] i wariancji $D^2(X)$=33
- Jeśli próbka modyfikowana borem (K) poddana została przesycaniu (H2) wtedy wytrzymałość będzie miała rozkład o średniej *E(X)*=577[Mpa] i wariancji $D^2(X)$=43
- Jeśli próbka modyfikowana borem (K) poddana została przesycaniu (H1) wtedy wytrzymałość będzie miała rozkład o średniej *E(X)*=546[Mpa] i wariancji $D^2(X)$=2187
- Jeśli próbka pochodząca z innego wytopu niż K poddana została przesycaniu (H2 lub H1) wtedy wytrzymałość będzie miała rozkład o średniej *E(X)*=600 [Mpa] i wariancji $D^2(X)$=325

Własności drzew

- Naturalna obsługa zmiennych mierzonych na różnych skalach pomiarowych
- Związki pomiędzy zmiennymi nie muszą być liniowe
- Rozkłady zmiennych nie muszą być normalne
- Jeśli spełnione są wymogi regresji wielorakiej to lepszy model daje regresja
- Drzewa nazywane – białą skrzynką – dobrze rozpoznany model i interpretacja

Własności drzew

- Niewrażliwość na zmienne bez znaczenia – mają niską ocenę ważności predyktorów
- Niewrażliwość na nadmierną korelację – jeśli dwie zmienne ze sobą skorelowane, jeden z predykatów nie wchodzi do drzewa
- Niewrażliwość na wartości odstające – podział w punkcie, nawet jeśli jakieś zmienne osiągają bardzo wysokie/niskie wartości
- Radzenie sobie z brakami danych – podziały zastępcze
- Naturalna interpretacja w postaci reguł
- Zastosowania: predykcja, budowa reguł, segmentacja rynku