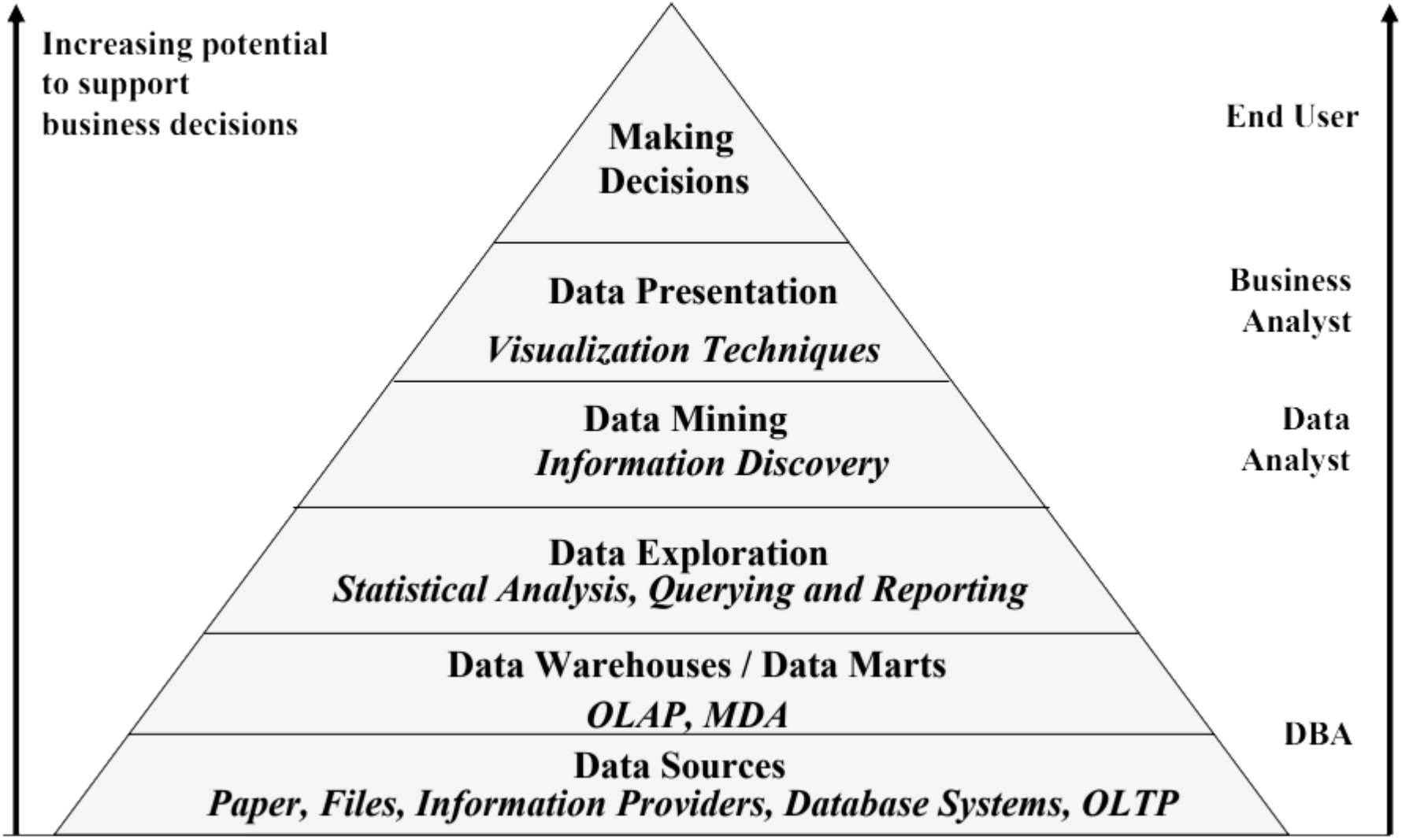
Indukcja reguł

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
3. J. Stefanowski – wykłady
4. Berthold, Borgelt, Höppner, Klawonn, *Guide to Intelligent Data Analysis*, Springer-Verlag London Limited 2010
5. Witten, Frank, *Data Mining Practical Machine Learning Tools and Techniques – WEKA*, Elsevier, San Francisco, 2005
6. *Data Science Central* - online resource for data practitioners

Krzysztof Regulski, WIMiIP, KISiM,
regulski@metal.agh.edu.pl

Data Mining and Business Intelligence

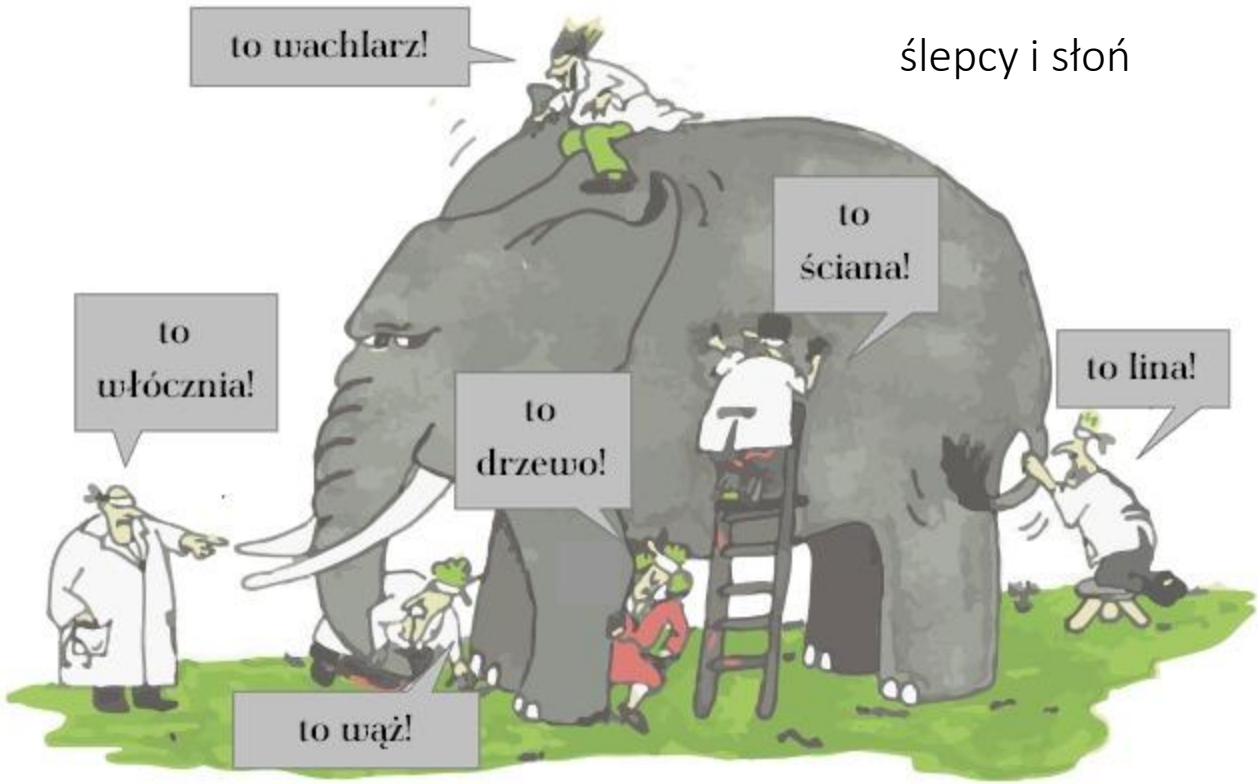


The story of Big Data, Data Science & Data Mining

datasciencecentral.com

data science: *"a set of fundamental principles that support and guide the principled extraction of information and knowledge from data."*

„data rich but information poor"

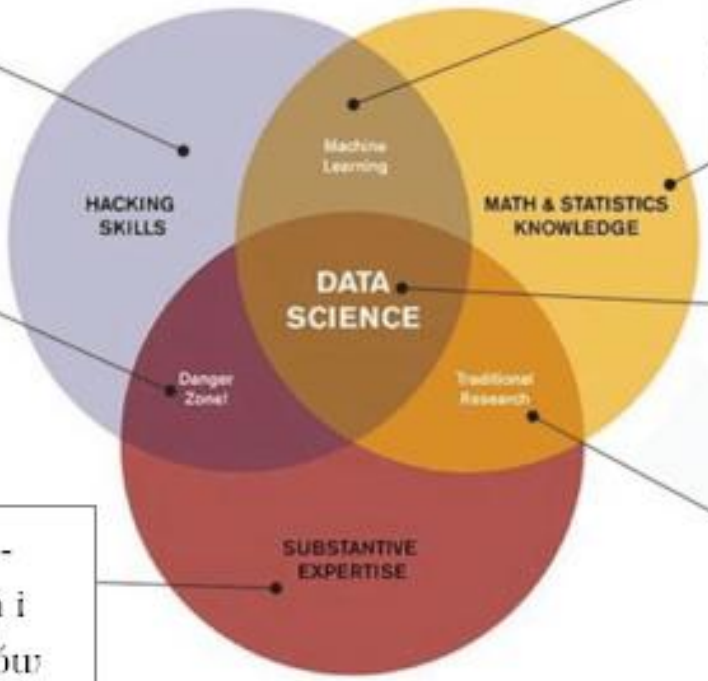


ślepcy i słoń

pozyskiwanie, czyszczenie i manipulowanie danymi

obszar ryzyka: informatyk + ekspert = ryzyko błędów metodycznych

wiedza merytoryczna - formułowanie wyzwań i interpretowanie wyników



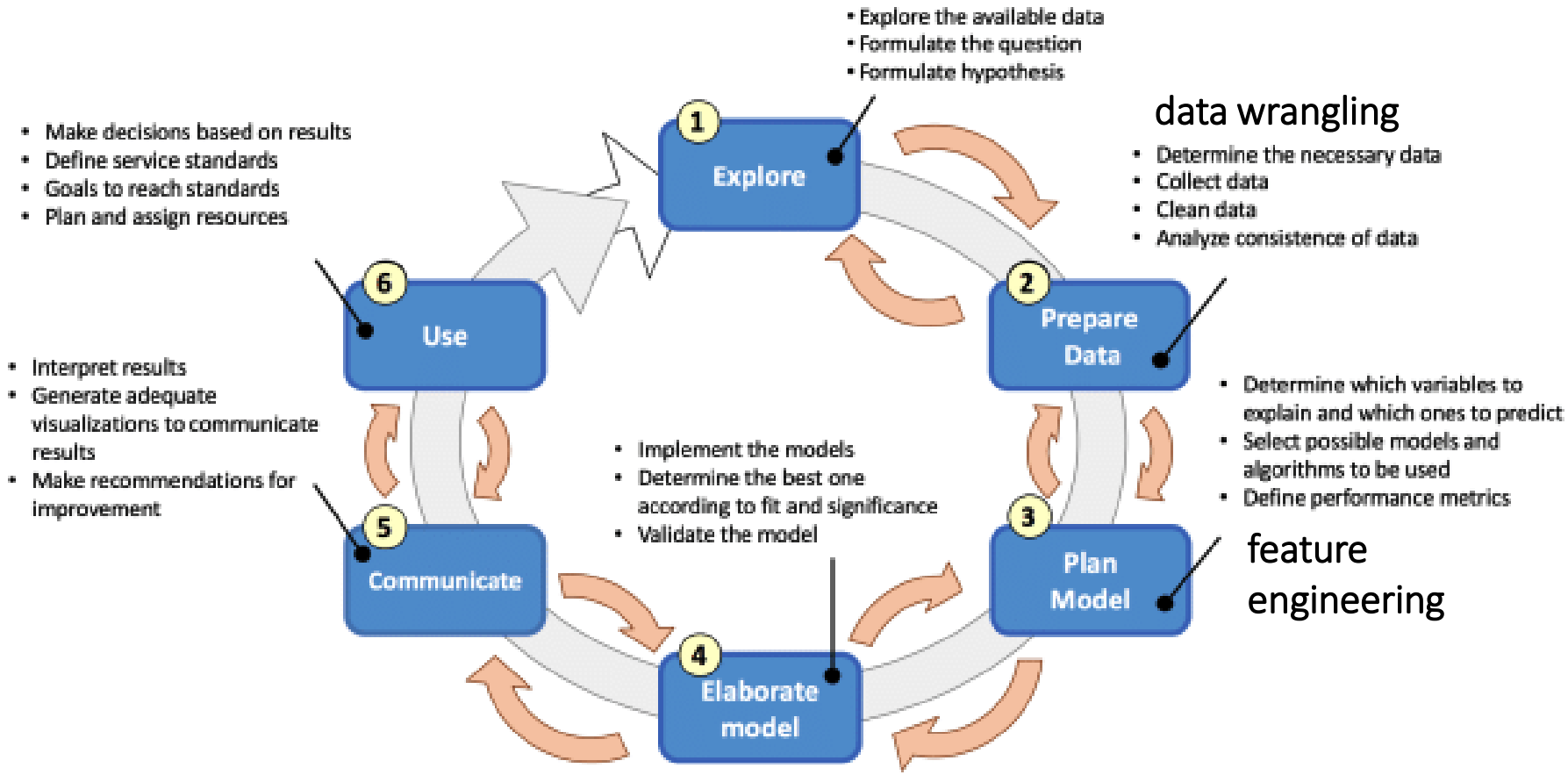
uczenie maszynowe = programowanie + statystyka

statystyka umożliwia dobór właściwych metod

data scientist: ekspert, statystyk i programista

tradycyjne badania i analiza danych pomiarowych

Knowledge Discovery from Data(KDD)



data mining - zadania

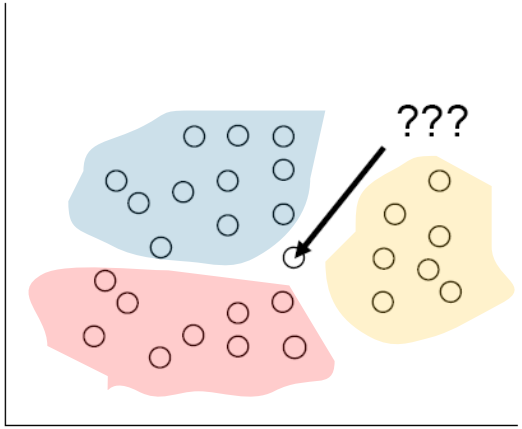
- **klasyfikacja:** przewidywanie wartości klasy na podstawie opisu
- **predykcja:** przewidywanie wartości ciągłej
- **analiza skupień:** poszukiwanie podobnych grup obiektów (skupień)
- **analiza asocjacji:** częste współwystępowanie sygnałów
- **analiza sekwencji:** analiza sekwencji sygnałów (obiektów) – analiza szeregów czasowych zmiennej jakościowej
- **charakterystyka:** tworzenie opisów grup
- **wizualizacja:** metody graficzne prezentacji wzorców
- **odkrywanie anomalii:** wykrywanie istotnych zmian (np. oszustw)

przede wszystkim prostota…

- proste algorytmy często wystarczają
- jest wiele przypadków prostych struktur danych:
  - jeden atrybut załatwia sprawę
  - wszystkie atrybuty mają podobny wpływ i są niezależne
  - logiczna struktura i mała liczba atrybutów odpowiednich dla drzew
  - zbiór kilku reguł logicznych
  - zależności pomiędzy grupami atrybutów
  - liniowa kombinacja atrybutów
  - wyraźne sąsiedztwo obiektów mierzone na podstawie odległości
  - wyraźne skupienia obiektów dla danych nieskategoryzowanych
  - zbiór obiektów dających się agregować
- skuteczność metod zależy od dziedziny/problemu badawczego

Klasyfikacja (1)

- 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 (2)

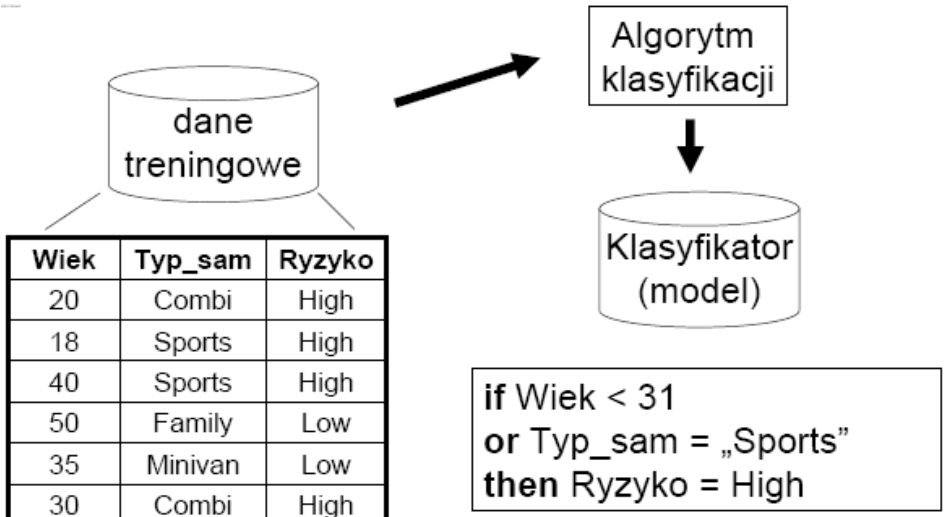
– **Dane wejściowe**

treningowy zbiór obserwacji będących listą wartości atrybutów opisowych (tzw. deskryptorów) i wybranego atrybutu decyzyjnego *(ang. class label attribute)*

– **Dane wyjściowe**

model (klasyfikator), **przydziela każdej krotce wartość atrybutu decyzyjnego** na podstawie wartości pozostałych atrybutów (deskryptorów, predyktorów)

Klasyfikacja – algorytm



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 sprawcą 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

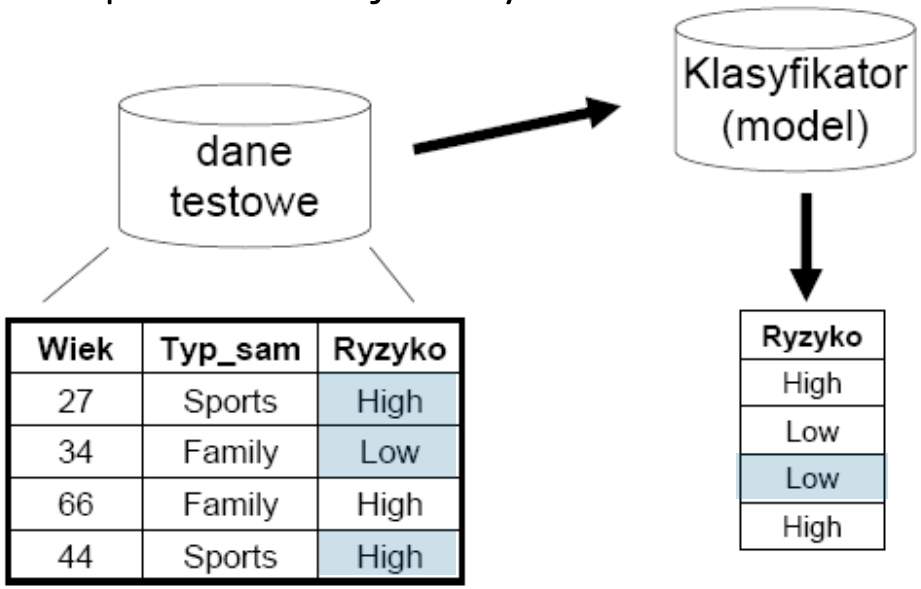
- Wynik klasyfikacji:
  - » Reguły klasyfikacyjne postaci *IF - THEN*
  - » Drzewa decyzyjne
- Istotną sprawą z punktu widzenia poprawności i efektywności modelu jest tzw. dokładność modelu.
- dla przykładów testowych, dla których znane są wartości atrybutu decyzyjnego, wartości te są porównywane z wartościami 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.



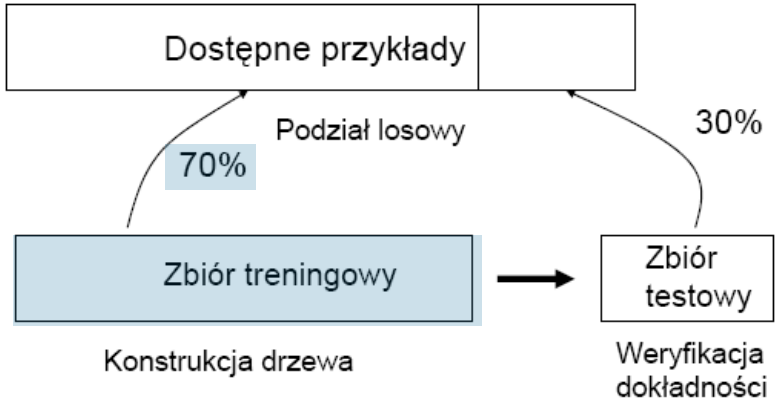
Wiek	Typ_sam	Ryzyko
27	Sports	High
34	Family	Low
66	Family	High
44	Sports	High

Ryzyko
High
Low
Low
High

Dokładność = 3/4 = 75%

Testowanie

Duży zbiór danych



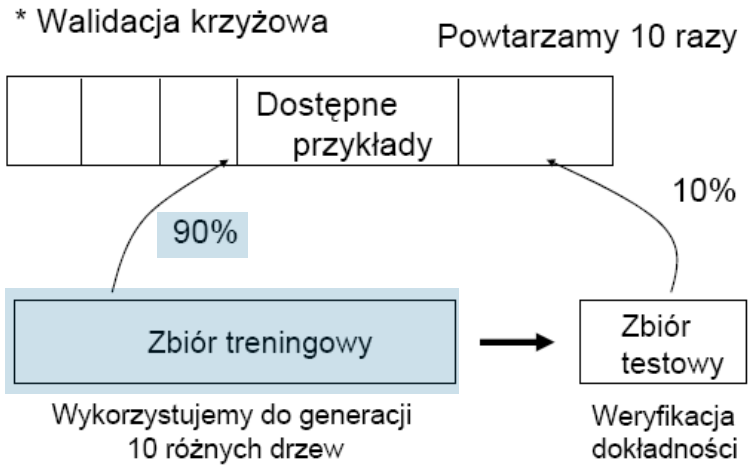
Mały zbiór danych

W przypadku zbioru przykładów o małej liczności stosujemy najczęściej metodę k-krotnej walidacji krzyżowej (tzw. kroswalidacji).

Początkowy zbiór przykładów jest losowo dzielony na k możliwie równych, wzajemnie niezależnych części S1, S2, ... , Sk.

Zbiór treningowy stanowi k-1 części, k-ta cześć stanowi zbiór testowy. Sam klasyfikator konstruujemy k-krotnie. W ten sposób otrzymujemy k-klasyfikatorów

Po wybraniu klasyfikatora, klasyfikator konstruuje się raz jeszcze w oparciu o cały dostępny zbiór przykładów



Sprawdzian krzyżowy

cross-validation

1. Dzieli się dane na *v* rozłącznych części (wybranych losowo, losowanie bez zwracania).
2. Dla ustalonego wstępnie *K* wykonuje się analizę, by znaleźć predykcję dla *v-tej* grupy danych (używając pozostałych *v-1* części danych jako przypadków "przykładowych").
3. Liczymy błąd predykcji. W przypadku regresji obliczamy sumę kwadratów reszt, przy klasyfikacji obliczamy dokładność, czyli procent przypadków zaklasyfikowanych poprawnie.
4. Na końcu *v* cykli uśredniamy błędy, otrzymując miarę jakości modelu.

   Powyższe powtarzamy dla różnych *K*, wybierając jako najlepsze to *K*, dla którego otrzymujemy najlepszą jakość modelu.

wielokrotne repróbkowanie

bootstrap

- metoda szacowania dokładności klasyfikatora dla mało licznego zbioru przykładów
- wykorzystuje losowanie ze zwracaniem
- tworzymy nowy zbiór losując n razy (z n-elementowego zbioru)
- niektóre przykłady będą się powtarzać w zbiorze treningowym, a inne przykłady w tym zbiorze nie wystąpią (dokładnie 0,368% przykładów nie zostanie wylosowanych)
- nowy zbiór obejmie 63,2% przypadków (*0.632 bootstrap*)
- niewylosowane przykłady utworzą zbiór testowy, który wykorzystujemy do oceny dokładności otrzymanego klasyfikatora.

$$\left(1-\frac{1}{n}\right)^n \approx e^{-1} = 0.368$$

Predykcja

- **klasyfikacja:** przewidywanie wartości klasy na podstawie opisu (wartości innych zmiennych)
- **predykcja:** przewidywanie wartości ciągłej, modelowanie funkcji ciągłych

Jeśli atrybut decyzyjny jest ciągły (numeryczny), mówimy o problemie predykcji / regresji.

Predykcja jest bardzo podobna do klasyfikacji. Jednakże celem predykcji jest zamodelowanie funkcji ciągłej, która by odwzorowywała wartości atrybutu decyzyjnego (regresja)

Rodzaje modeli klasyfikacyjnych:

- k-NN - k-najbliższe sąsiedztwo
  (wartościowanie leniwe, lazy evaluation)
- Klasyfikatory Bayes'owskie
  (wartościowanie zachłanne, eager evaluation) - generative model approaches
- Klasyfikacja poprzez indukcję drzew decyzyjnych
- Klasyfikatory liniowe (LDA, FLA, etc.)
- Discriminative Modelling Approaches
  (Linear Classifier, Logistics Regression, SVM)
- SVM – (Support Vector Machine) - Metoda wektorów nośnych
- indukcja reguł, zbiory przybliżone, reguły asocjacyjne
- Sieci Neuronowe, neuro-fuzzy
- Analiza statystyczna, Metaheurystyki
  (np. algorytmy genetyczne)
- i inne…

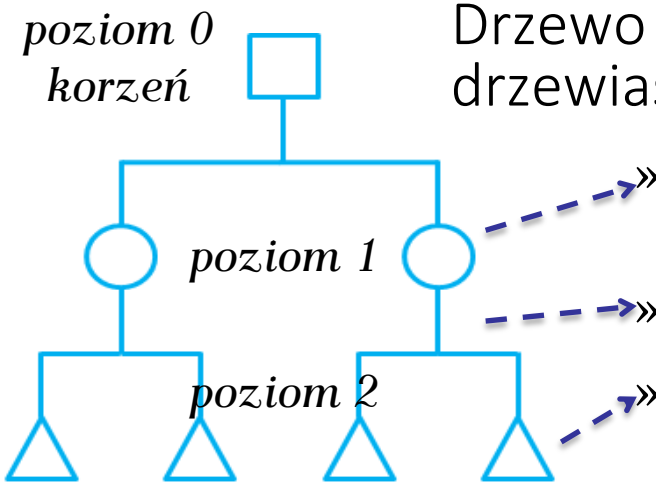
Indukcja drzew klasyfikacyjnych

zmienna zależna: jakościowa

DTs (decision trees)

- rodzina algorytmów uczenia maszynowego
- uczenie nadzorowane (*supervised*), z nauczycielem
- dane są opisane kategoriami (etykietowane, *labelled*)
- uczenie wykorzystywane m.in.
    - w autonomicznych pojazdach do rozpoznawania obiektów,
    - w medycynie,
    - w ocenie wartości klientów,
    - chatterbotach,
    - itd.

Klasyfikacja poprzez indukcję drzew decyzyjnych



Drzewo decyzyjne jest grafem o strukturze drzewiastej, gdzie

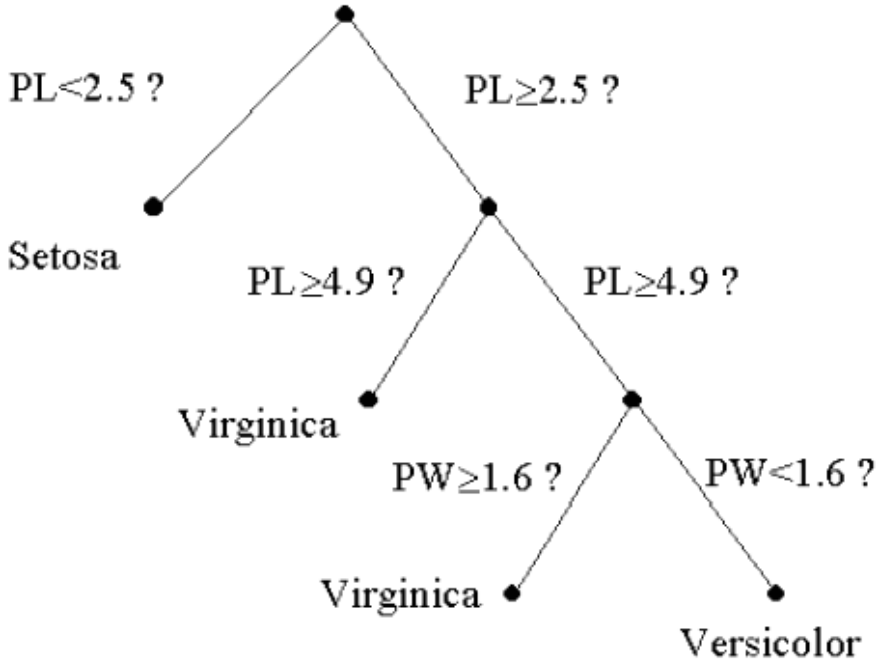
- » każdy węzeł reprezentuje test na atrybucie (podział),
- » każdy łuk (gałąź) reprezentuje wynik testu,
- » każdy liść reprezentuje pojedynczą klasę lub rozkład wartości klas

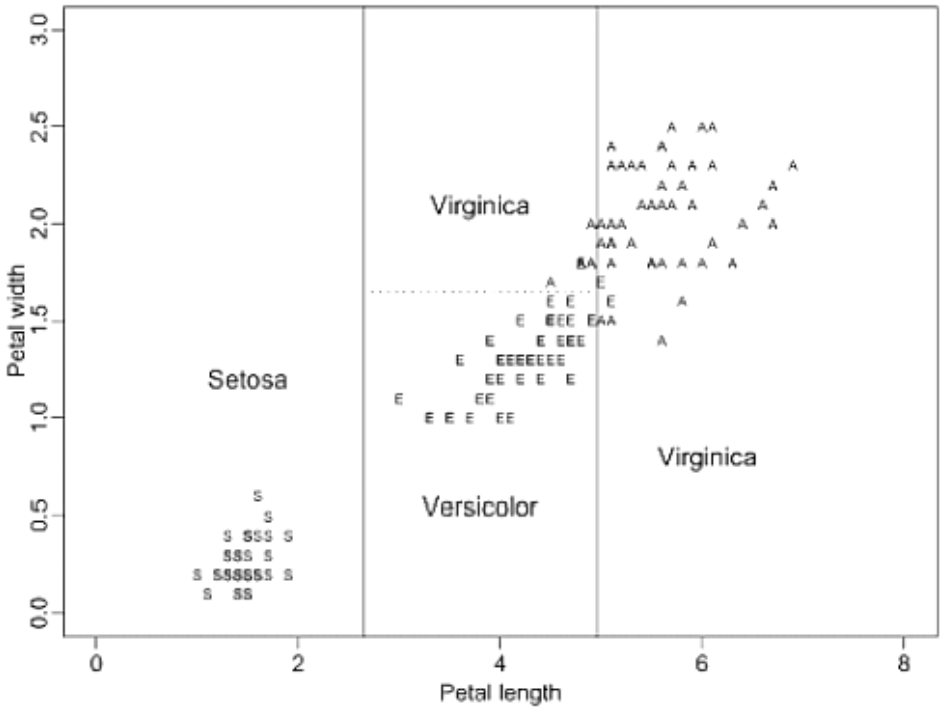
Drzewo decyzyjne dzieli zbiór treningowy na partycje do momentu, w którym każda partycja zawiera dane należące do jednej klasy, lub, gdy w ramach partycji dominują dane należące do jednej klasy

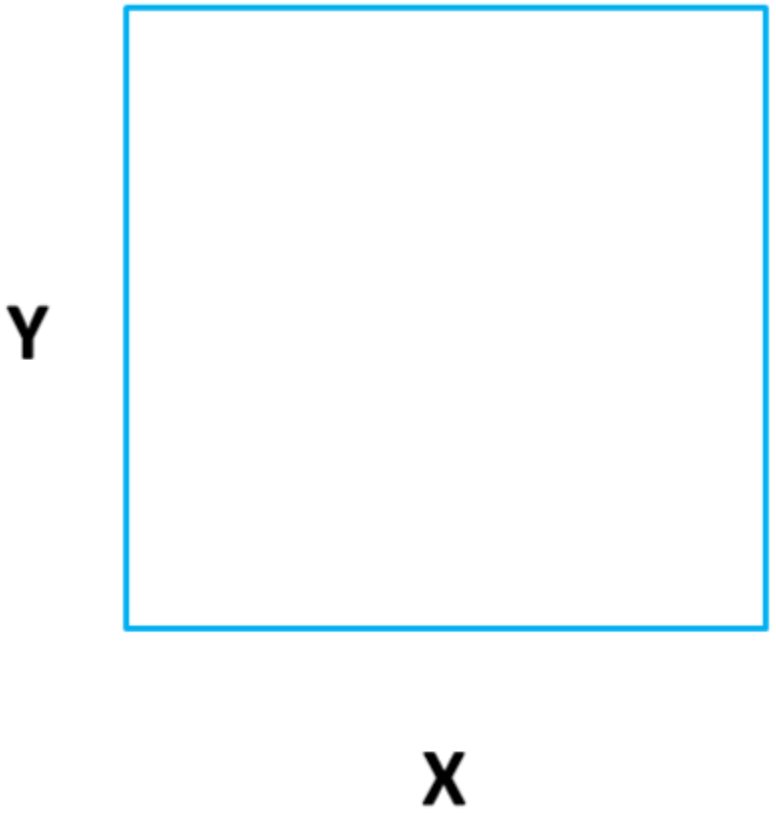
Każdy wierzchołek wewnętrzny (węzeł) drzewa zawiera tzw. **punkt podziału** *(ang. split point),* którym jest test na atrybucie (atrybutach), który dzieli zbiór danych na partycje

indukcja drzew

- *divide-and-conquer*
  1: wybierz atrybut podziału korzenia (pełnego zbioru)
  utwórz gałąź dla każdej wartości atrybutu
  2: podziel zbiór na partycje
  dla każdej gałęzi
  3: powtarzaj rekursywnie dla każdej gałęzi
  używaj tylko instancji należących do partycji
- przerwij, jeśli wszystkie instancje należą do jednej klasy


PL<2.5 ?
PL≥2.5 ?
Setosa
PL≥4.9 ?
PL≥4.9 ?
Virginica
PW≥1.6 ?
PW<1.6 ?
Virginica
Versicolor

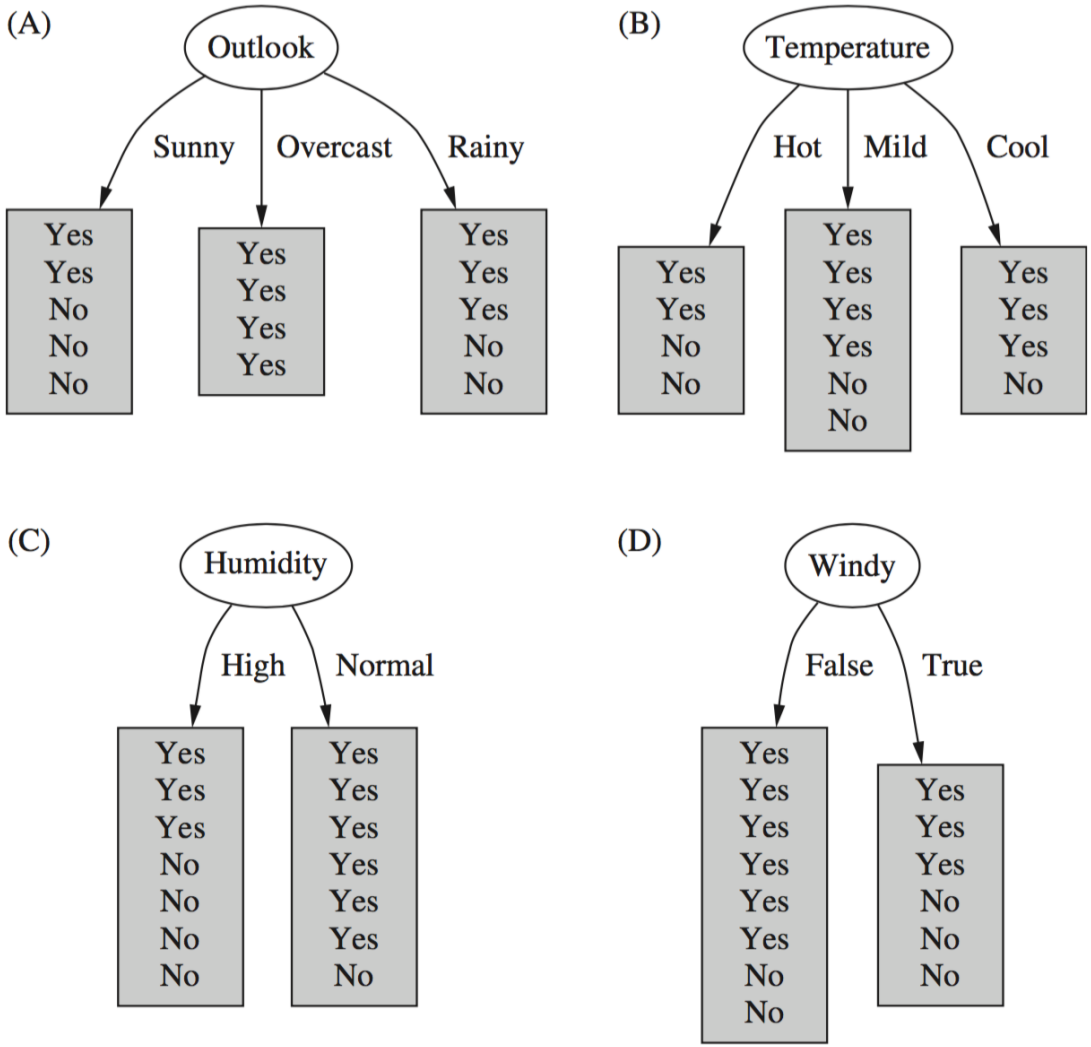

Petal width
Petal length
Setosa
Virginica
Versicolor
Virginica
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
2
4
6
8


Y
X

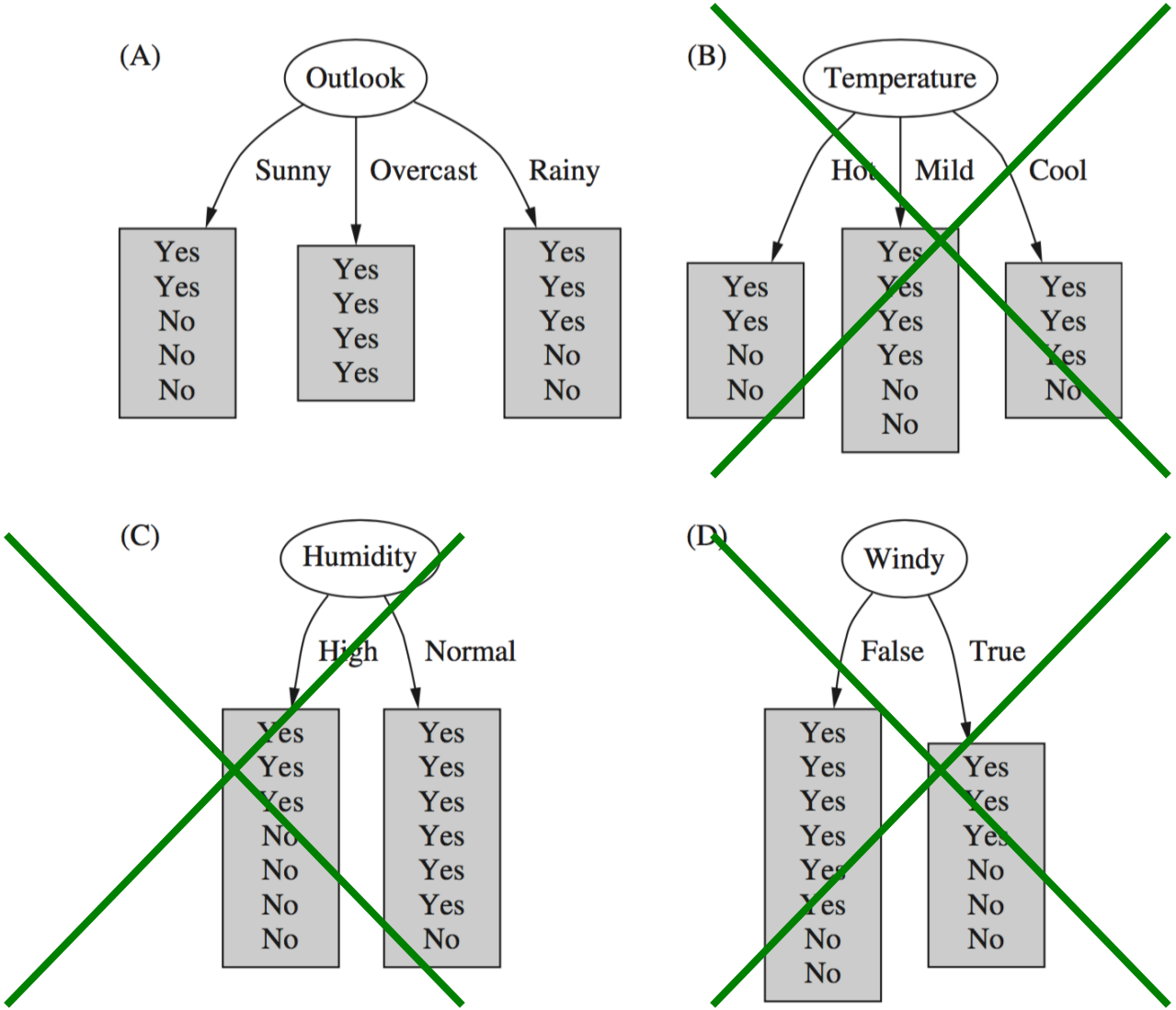
Czy dziś grać w golfa?

zachmurzenie	temperatura	wilgotność	wiatr	decyzja
słońce	gorąco	wysoka	słaby	nie
słońce	gorąco	wysoka	silny	nie
pochmurno	gorąco	wysoka	słaby	tak
deszcz	średnio	wysoka	słaby	tak
deszcz	chłodno	normalna	słaby	tak
deszcz	chłodno	normalna	silny	nie
pochmurno	chłodno	normalna	silny	tak
słońce	średnio	wysoka	słaby	nie
słońce	chłodno	normalna	słaby	tak
deszcz	średnio	normalna	słaby	tak
słońce	średnio	normalna	silny	tak
pochmurno	średnio	wysoka	silny	tak
pochmurno	gorąco	normalna	słaby	tak
deszcz	średnio	wysoka	silny	nie

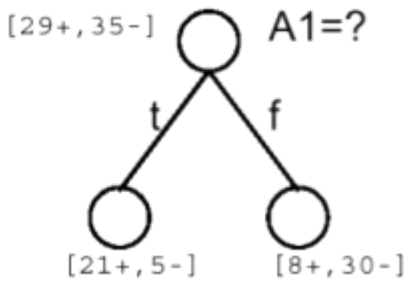
wybór predyktora (1)

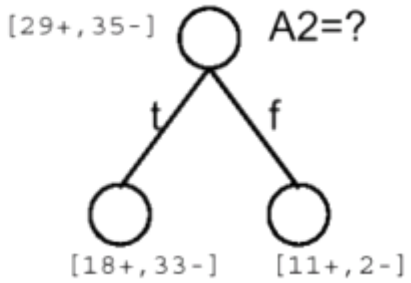


wybór predyktora (2)



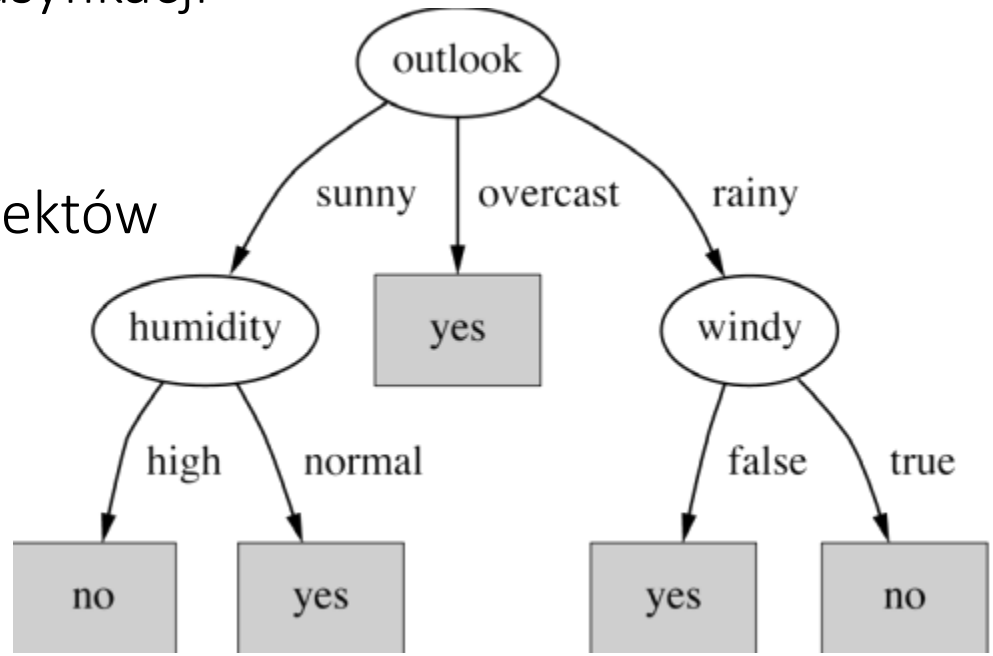
wybór predyktora (3)



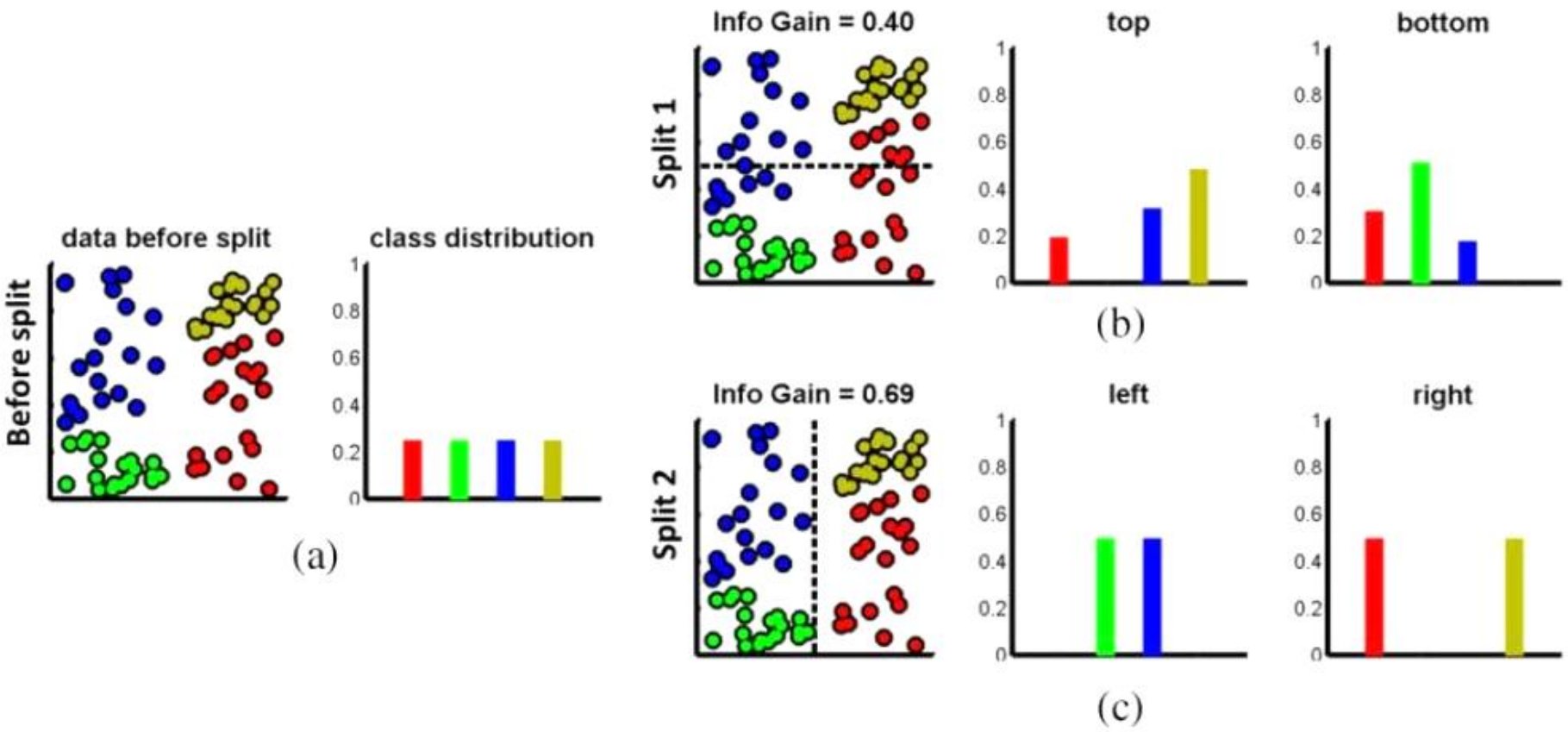


miara (nie)czystości:
- information gain,
- entropy
- Gini Index
- χ^2

- atrybut ma być użyteczny w klasyfikacji
- potrzebna jest miara jakości (im)purity function
- mierzymy stopień separacji obiektów względem klas
- wybieramy atrybut dający „najczystszy” klasowo podział i najmniejsze drzewo



wybór predyktora (4)



Kryteria oceny podziału

Indeks Gini (algorytmy CART, SPRINT)

Wybieramy atrybut, który minimalizuje indeks Gini

gdzie:

$$gini(S) = 1 - \sum p_j^2$$

- **S** – zbiór przykładów należących do n klas
- p_j – względna częstość występowania klasy j w S

Zysk informacyjny (algorytmy ID3, C4.5)

Wybieramy atrybut, który maksymalizuje redukcję entropii

Entropia jest miarą stopnia nieuporządkowania. Im mniejsza wartość entropii, tym większa „czystość" podziału zbioru S na partycje

$$E(A_1, A_2, ...A_v) = \sum_{j=1}^{v} \frac{(s_{1j} + s_{2j} + ... + s_{mj})}{s} I(s_{1j}, s_{2j}, ..., s_{mj})$$

Indeks korelacji $\chi 2$ (algorytm CHAID)

Mierzymy korelację pomiędzy każdym atrybutem i każdą klasą (wartością atrybutu decyzyjnego)

Wybieramy atrybut o maksymalnej korelacji

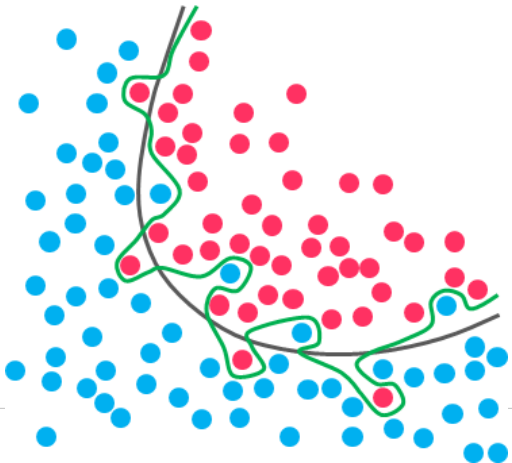
Obcinanie drzewa (1)

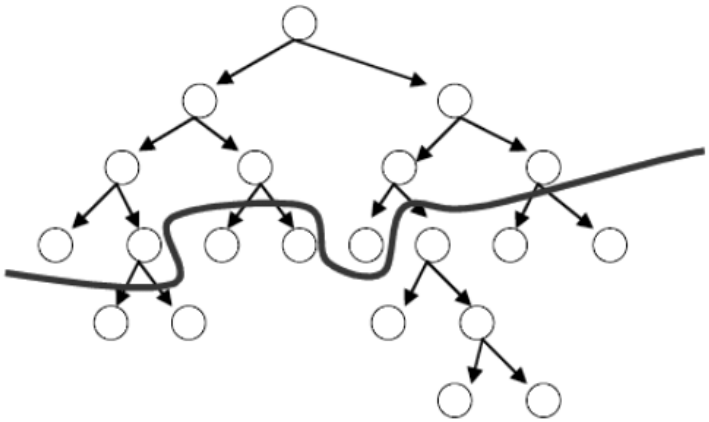
Algorytmy indukcji drzew dążą do możliwie najlepszej klasyfikacji, co prowadzi do przeuczenia (szum i punkty osobliwe)

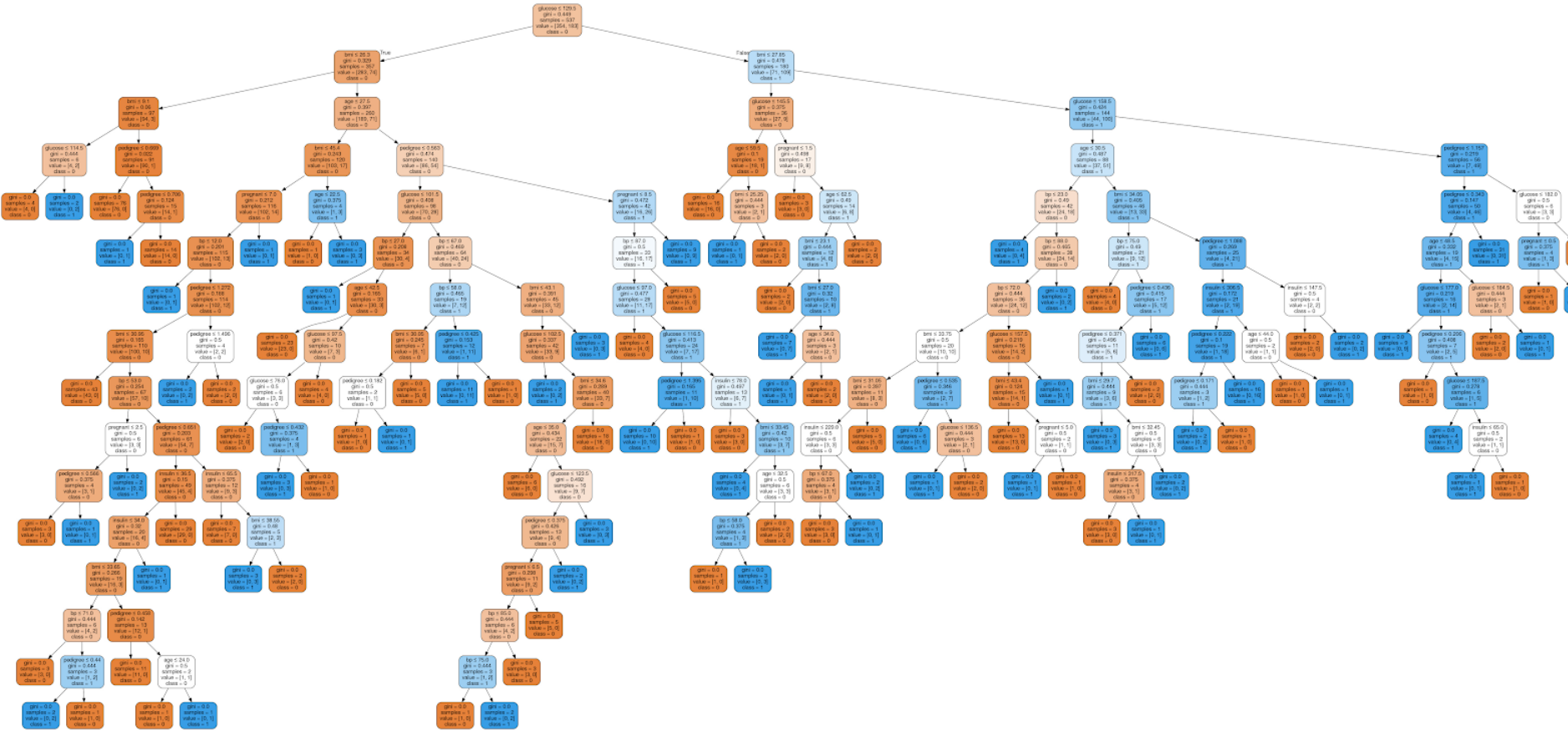
Przycinanie drzew decyzyjnych - usuwanie mało wiarygodnych gałęzi

- » poprawia efektywność klasyfikacji
- » poprawia zdolność klasyfikatora do klasyfikacji nowych przypadków

Metody przycinania drzew decyzyjnych - bazują najczęściej na miarach statystycznych







Przykład drzewa bez przycięcia (DataCamp)

Obcinanie drzewa (2)

Dwa podejścia do problemu przycinania drzew decyzyjnych:

– wstępne przycinanie drzewa decyzyjnego (*prepruning*)

» polega na przycięciu drzewa przez wcześniejsze zatrzymanie procedury konstrukcji drzewa. Wprowadzamy warunek stopu, który wstrzymuje dalsze dzielenie zbioru treningowego na partycje. Przykładowym warunkiem stopu jest przyjęcie minimalnej liczby elementów należących do partycji, która podlega dzieleniu.

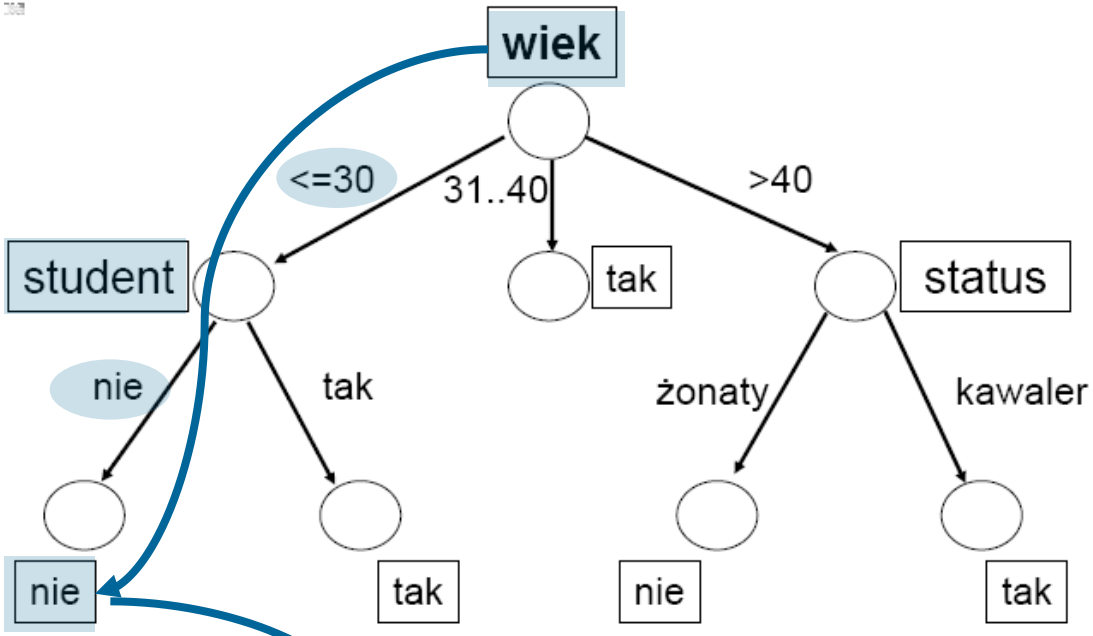
– przycinanie drzewa po zakończeniu konstrukcji drzewa (*postpruning*)

» bazuje na miarach statystycznych

ekstrakcja reguł (1)

- 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
- każda gałąź tworzy poprzednik (przesłanki) reguły klasyfikacyjnej: koniunkcja par *<atrybut, wartość>*
- każdy liść tworzy następnik (konkluzję) reguły

Ekstrakcja reguł klasyfikacyjnych z drzew decyzyjnych (2)

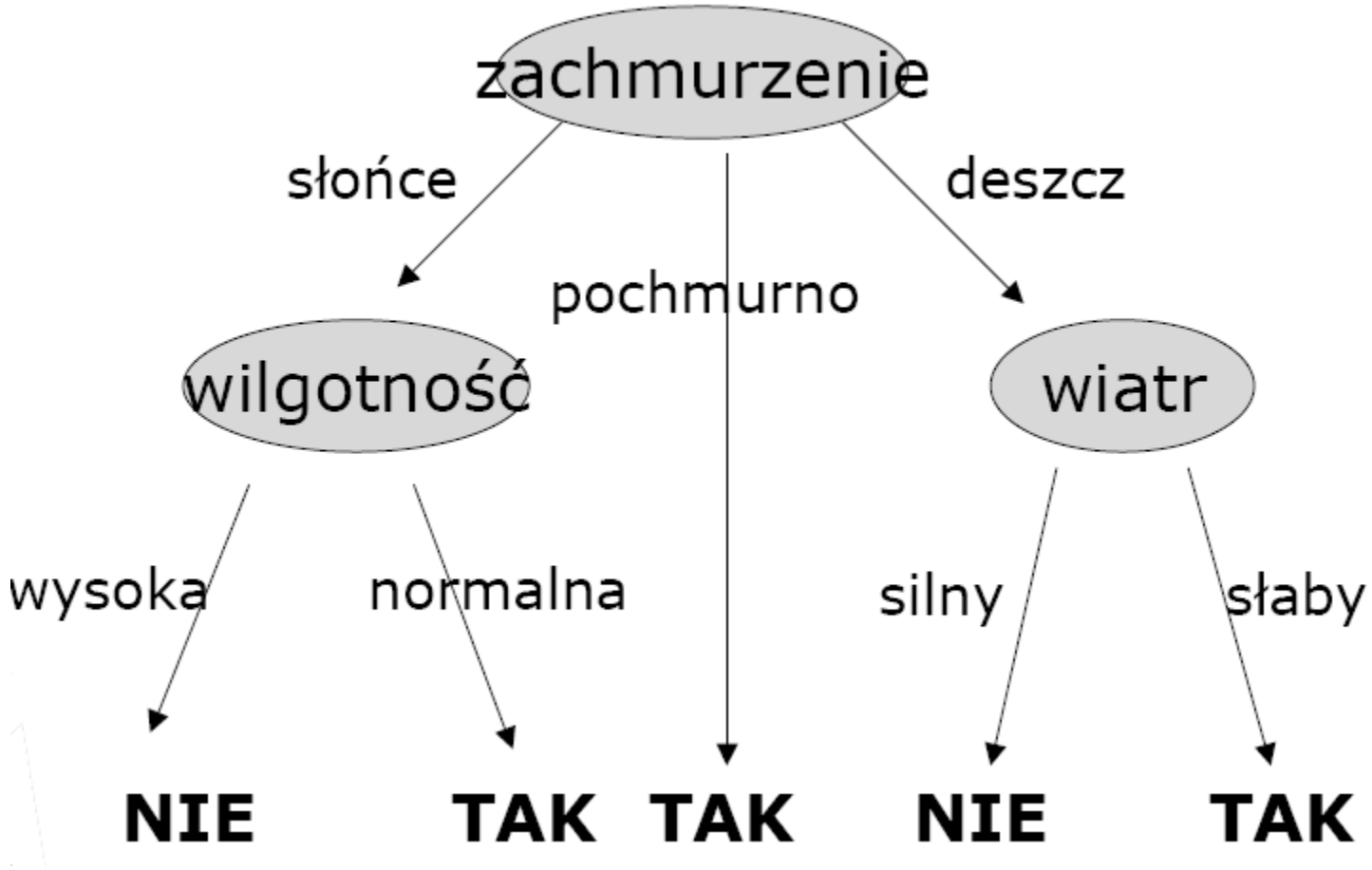


Drzewo decyzyjne można przedstawić w postaci następującego zbioru reguł klasyfikacyjnych:

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’

Klasyfikator



Drzewa – pojęcia podstawowe

- Jeśli zmienna zależna jest wyrażona na skalach słabych (jakościowych) to mówimy o drzewach **klasyfikacyjnych**,
- jeśli na mocnych (ilościowych), to o drzewach **regresyjnych**.
- Skala zmiennych objaśniających nie ma znaczenia.

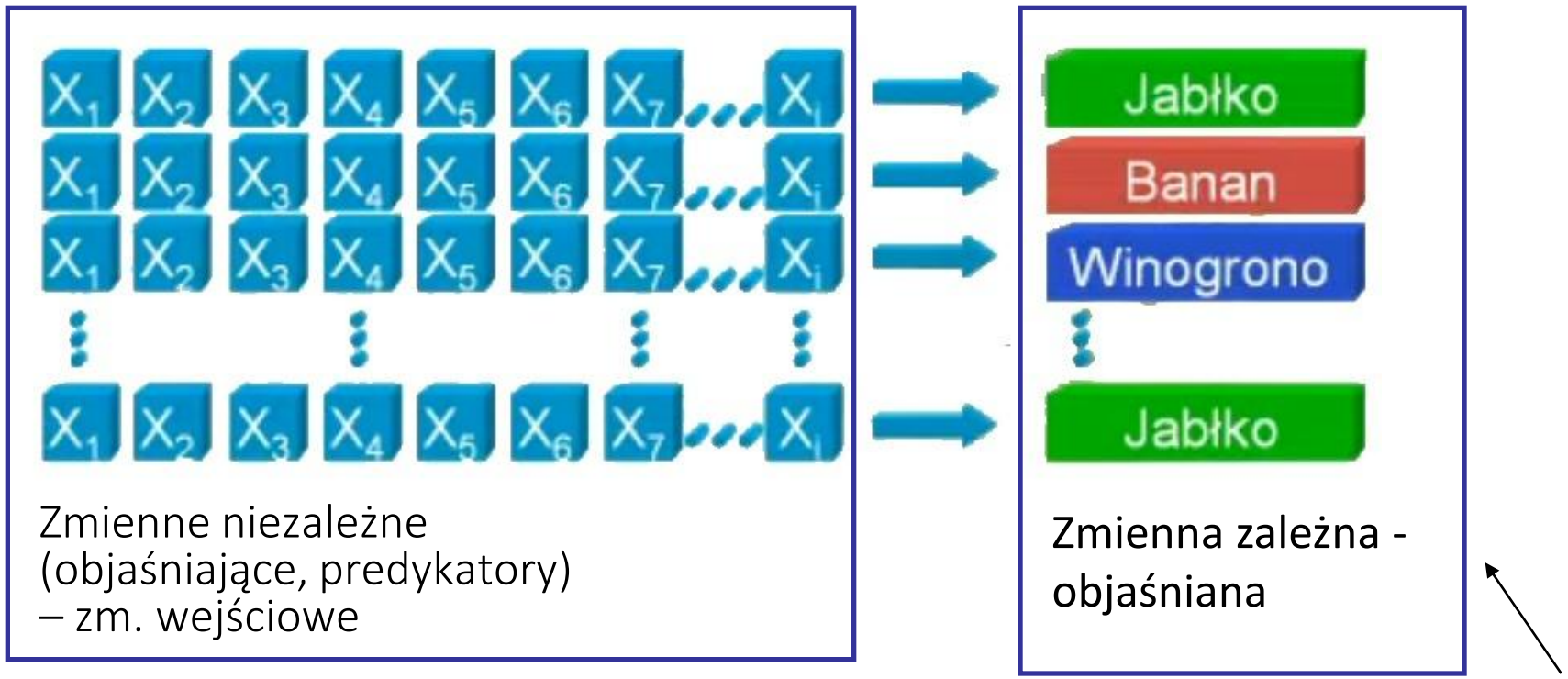
Węzły i liście

– Drzewem binarnym jest drzewo, w którym z każdego węzła wychodzą dwie krawędzie

– Liściem (węzłem końcowym) nazywamy węzeł, z którego nie wychodzą żadne krawędzie

– Wielkość drzewa to liczba liści, a głębokość drzewa to długość najdłuższej drogi między korzeniem a liściem (liczba krawędzi między tymi dwoma węzłami)

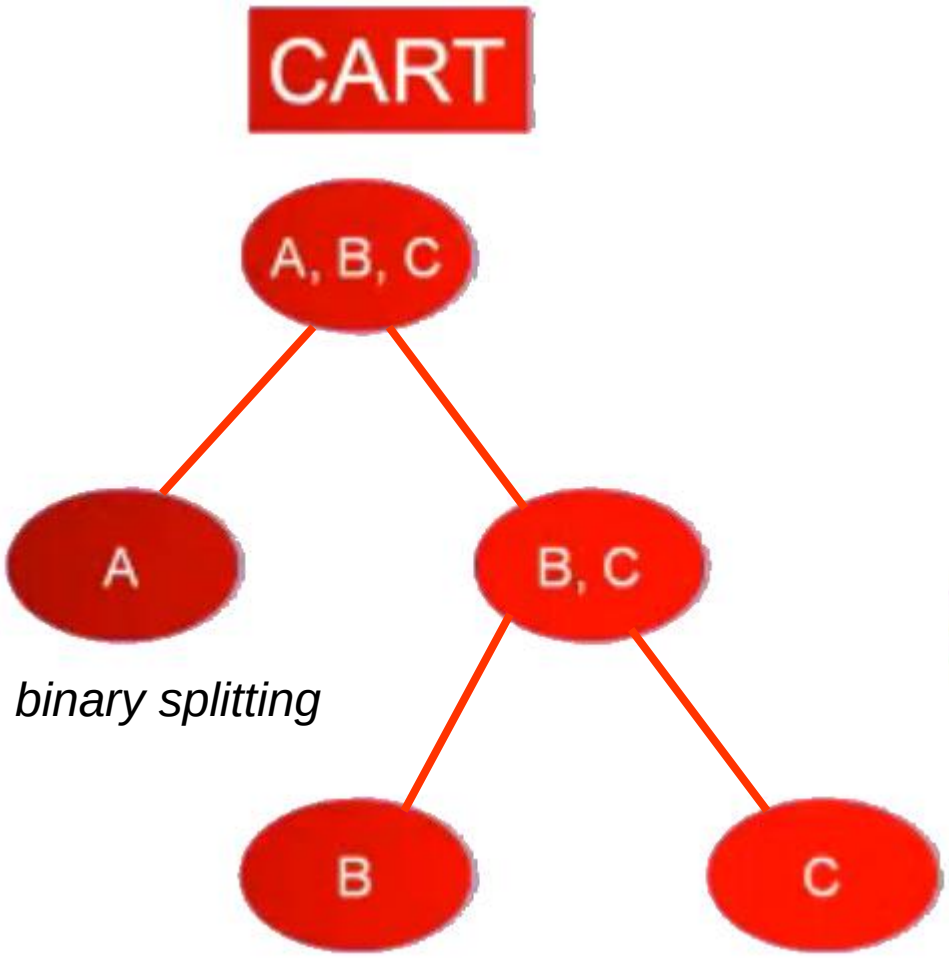
Zmienne w analizie



CART (C&RT) i CHAID

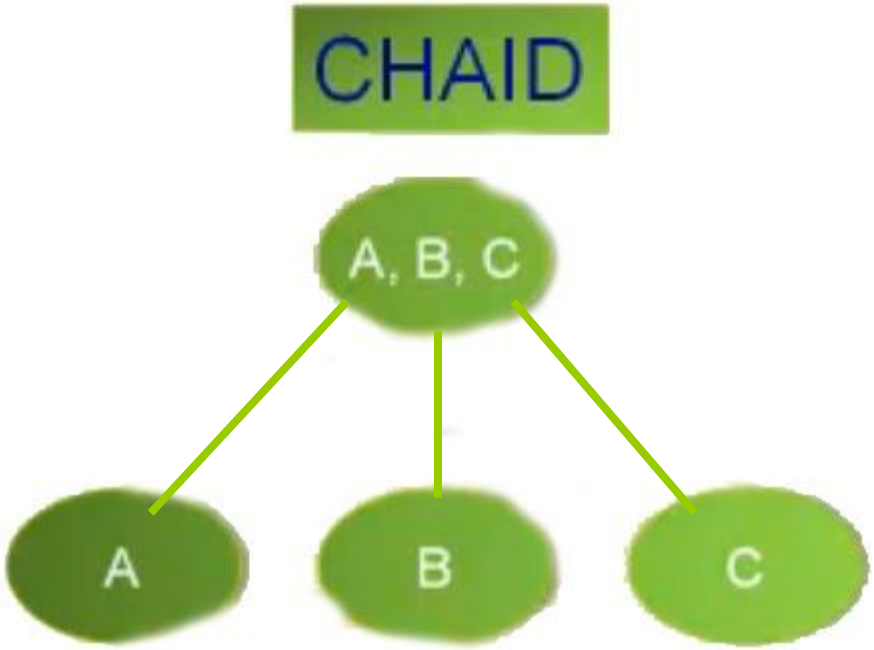
Classification & Regression Trees

C&RT



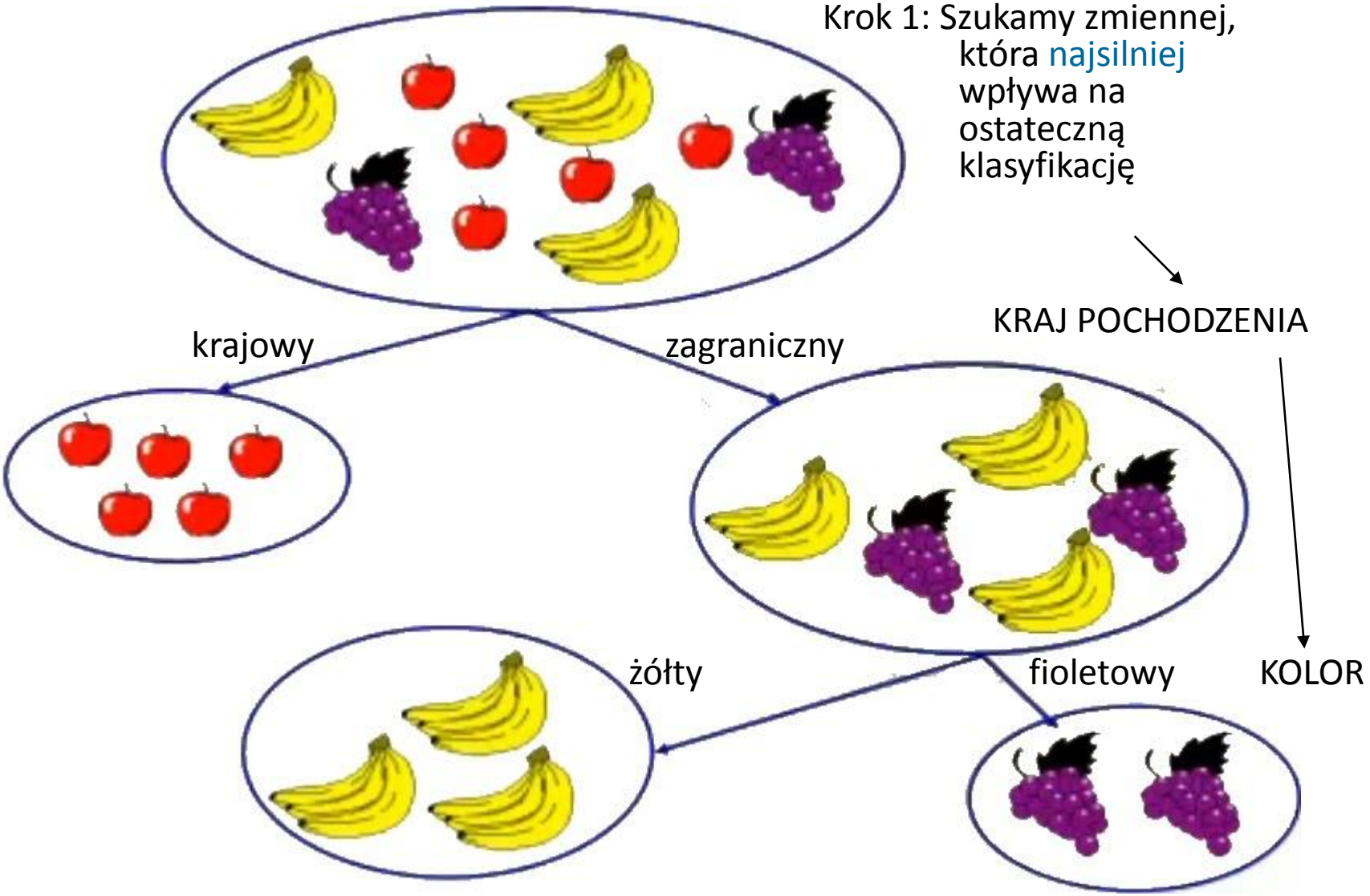
binary splitting

Chi-square Automatic Interaction Detection



multiway splitting

Drzewa klasyfikacyjne



C&RT

- Wykorzystuje każdą kombinację zmiennych ciągłych i kategorialnych (jakościowych)
- Wybiera najlepszy podział
- W kolejnych podziałach mogą być wykorzystywane te same zmienne
- Metoda niewrażliwa na obserwacje odstające
- Obsługuje zbiory obserwacji o złożonej strukturze

Przykład: segmentacja pod kątem dochodów

- Cel: segmentacja ze względu na dochód, charakterystyka segmentów
- Dane zawierają 32 tys przypadków
- Każdy przypadek reprezentuje jedną osobę
- Każda osoba opisana jest przez 11 cech demograficznych oraz zmienną dochód

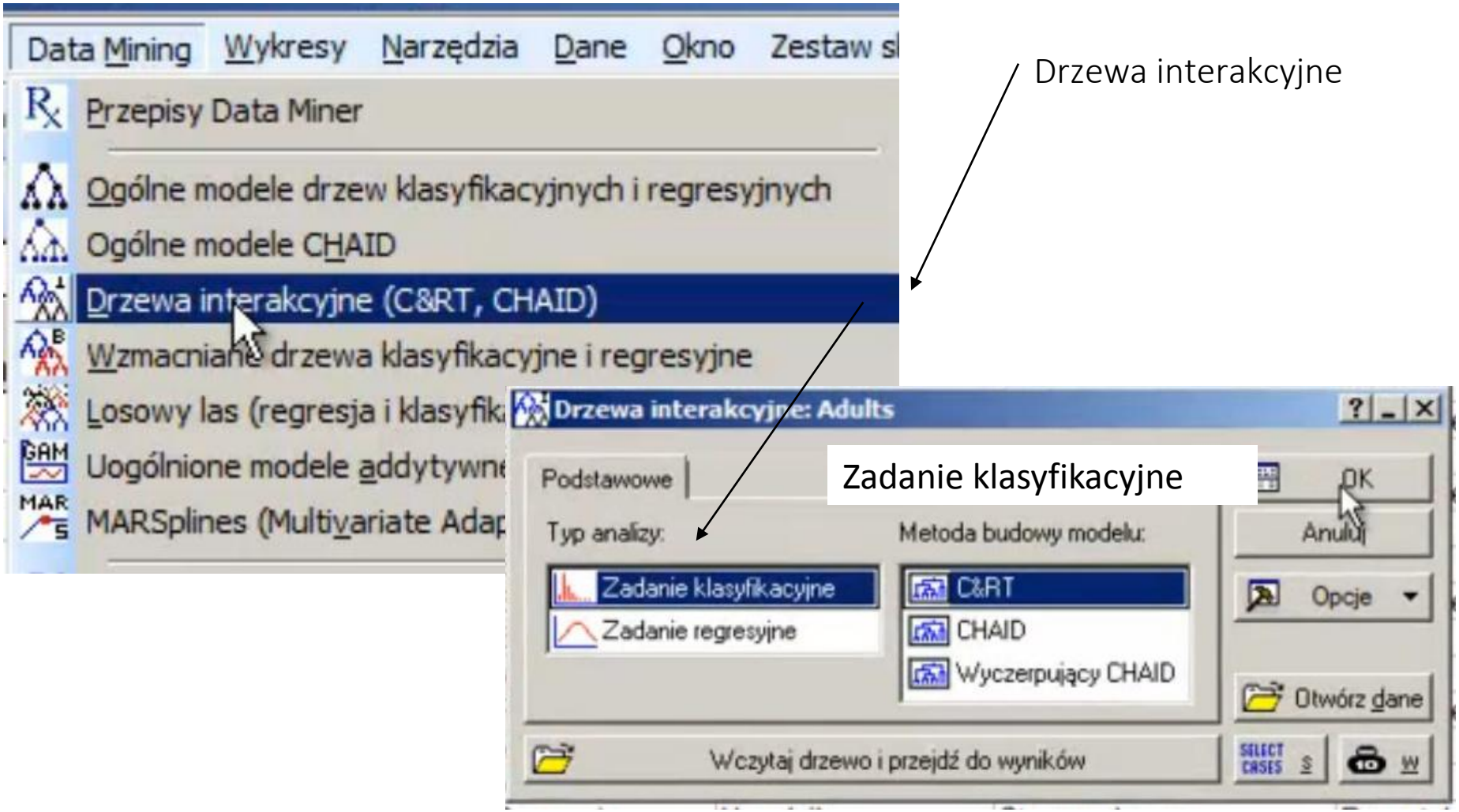
adults (11 zmn. * 32561 prz.)

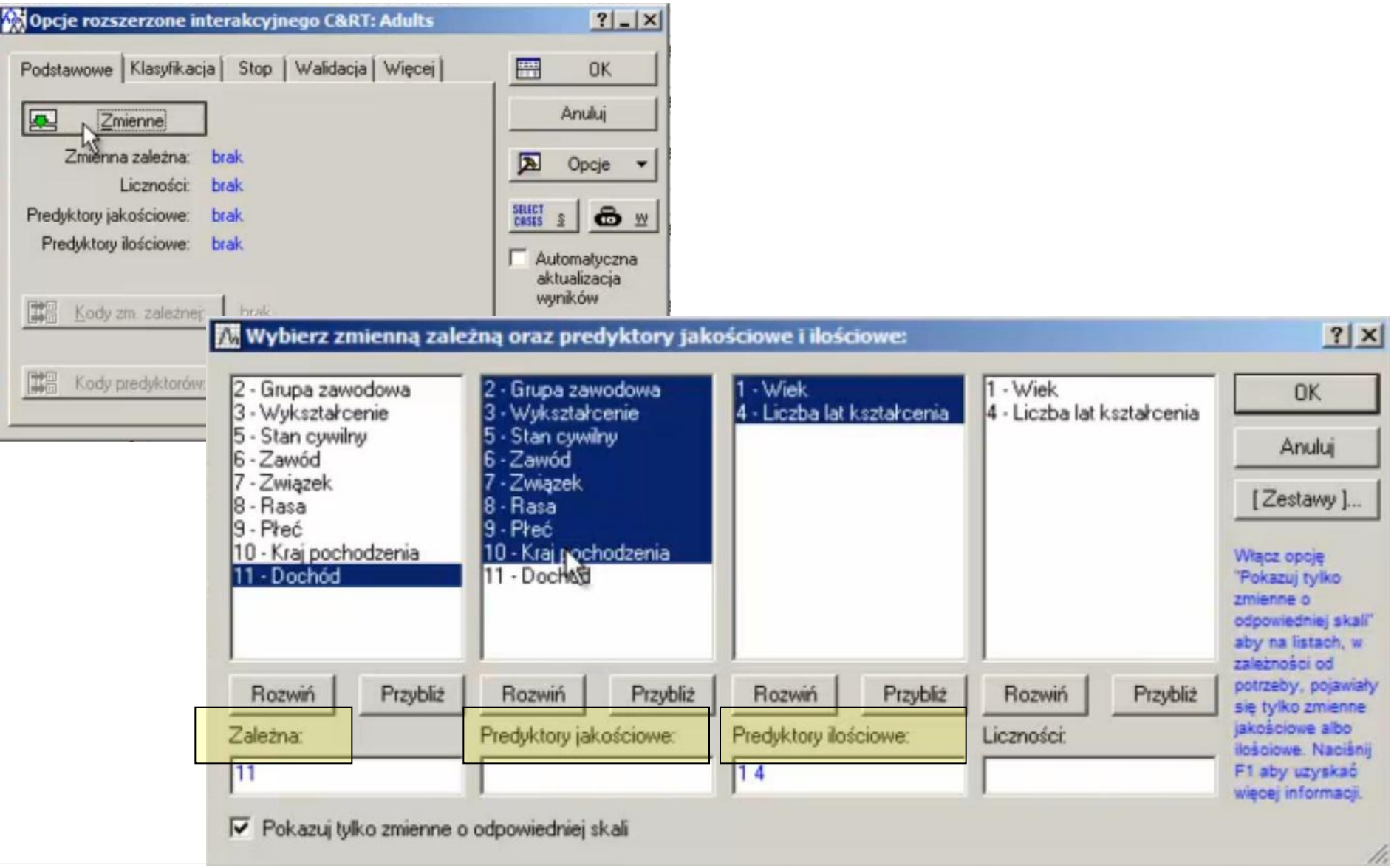
	1 Wiek	2 Grupa zawodowa	3 Wykształcenie	4 Liczba lat kształcenia	5 Stan cywilny	6 Zawód	7 Związek	8 Rasa	9 Płeć
1	25	Firma prywatna	College	10	Rozwiedziona/ny	Kierownik/Dyrektor	Stanu wolnego	Biała	Kobieta
2	39	Firma prywatna	Licencjat	12	Separacja	Mechanik	Poza rodziną	Biała	Mężczyzna
3	45	Firma prywatna	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 federalna	College	10	Separacja	Urzędnik	Stanu wolnego	Pozostałe	Kobieta
18	26	Firma prywatna	Podstawowe	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	Podstawowe	8	Stanu wolnego	Usługi ochroniarskie	Wychowująca(cy) dziecko	Pozostałe	Mężczyzna

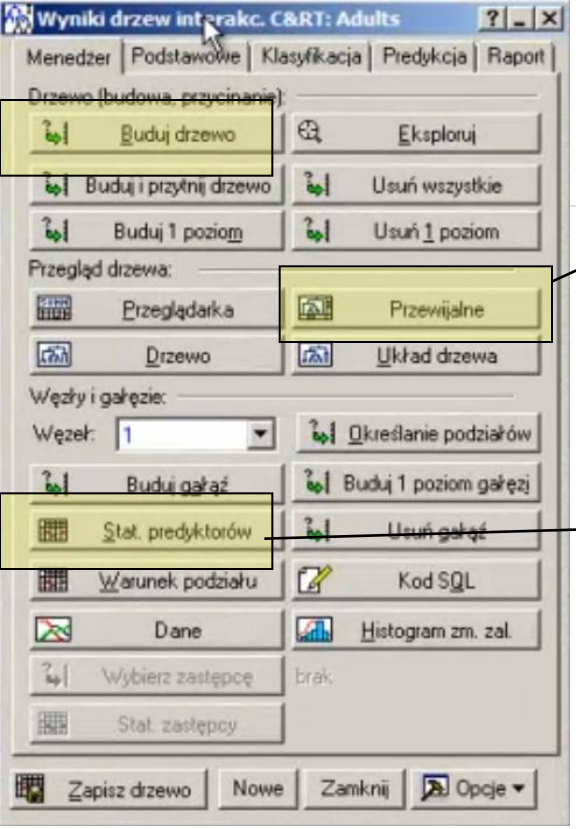
Dochód skategoryzowany do 2 klas:

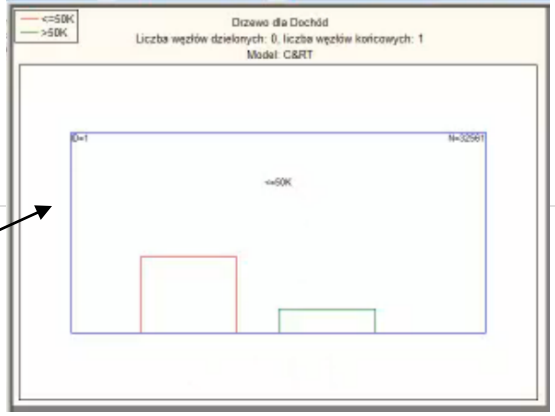
<=50K – poniżej 50 000

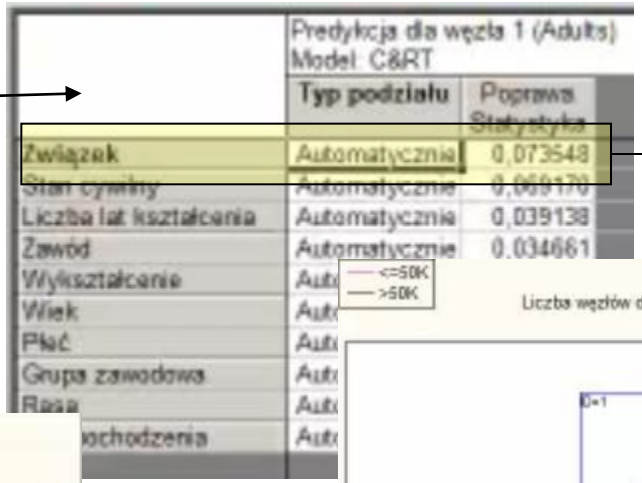
>50K – powyżej 50 000

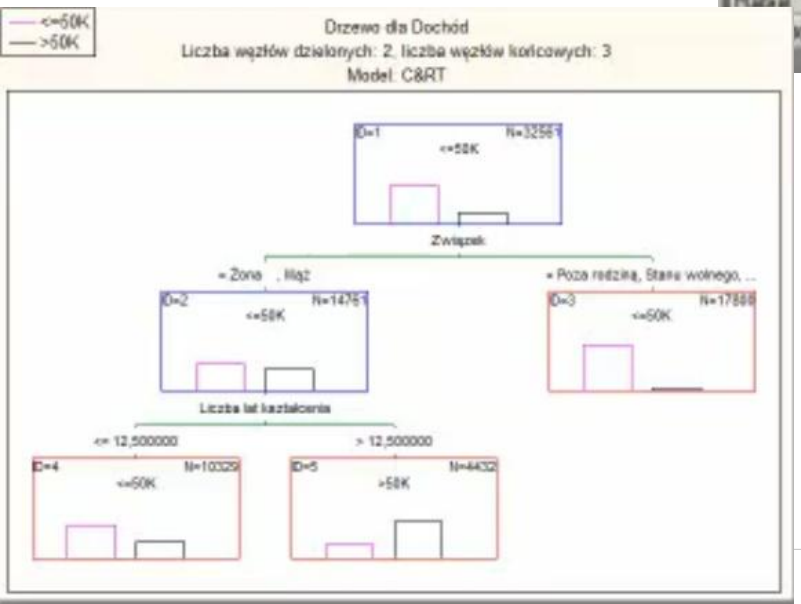

Data Mining
Wykresy
Narzędzia
Dane
Okno
Przepisy Data Miner
Ogólne modele drzew klasyfikacyjnych i regresyjnych
Ogólne modele CHAID
Drzewa interakcyjne (C&RT, CHAID)
Wzmacniane drzewa klasyfikacyjne i regresyjne
Losowy las (regresja i klasyfik
Uogólnione modele addytywne
MARSplines (Multivariate Adap
Drzewa interakcyjne
Drzewa interakcyjne: Adults
Podstawowe
Zadanie klasyfikacyjne
Typ analizy:
Zadanie regresyjne
Metoda budowy modelu:
C&RT
CHAID
Wyczerpujący CHAID
OK
Anuluj
Opcje
Otwórz dane
Wczytaj drzewo i przejdź do wyników

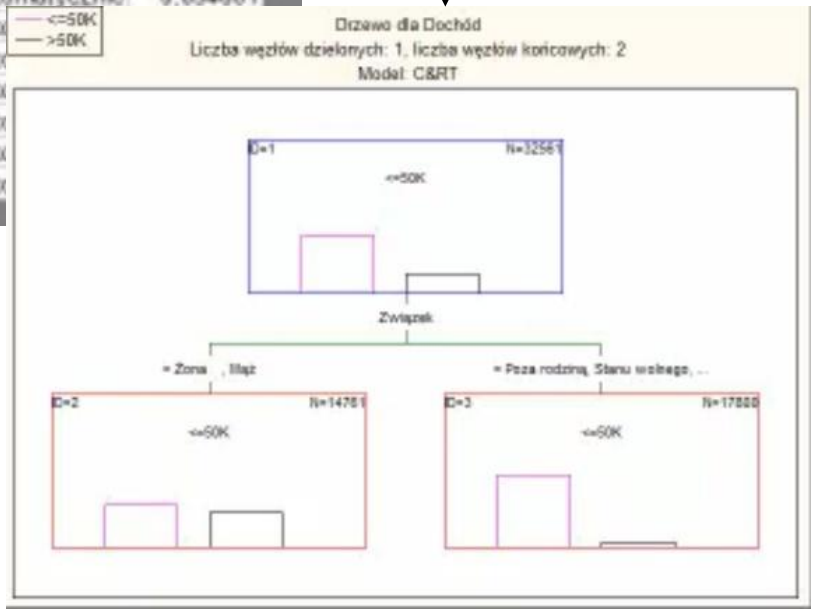

Opcje rozszerzone interakcyjnego C&RT: Adults
Podstawowe
Klasyfikacja
Stop
Walidacja
Więcej
OK
Anuluj
Zmienne
Zmienna zależna: brak
Liczności: brak
Predyktory jakościowe: brak
Predyktory ilościowe: brak
Opcje
Automatyczna aktualizacja wyników
Kody zm. zależnej:
Kody predyktorów:
Wybierz zmienną zależną oraz predyktory jakościowe i ilościowe:
2 - Grupa zawodowa
3 - Wykształcenie
5 - Stan cywilny
6 - Zawód
7 - Związek
8 - Rasa
9 - Płeć
10 - Kraj pochodzenia
11 - Dochód
1 - Wiek
4 - Liczba lat kształcenia
[Zestawy]...
Włącz opcję "Pokazuj tylko zmienne o odpowiedniej skali" aby na listach, w zależności od potrzeby, pojawiały się tylko zmienne jakościowe albo ilościowe. Naciśnij F1 aby uzyskać więcej informacji.
Rozwiń
Przybliż
Zależna:
Predyktory jakościowe:
Predyktory ilościowe:
Liczności:
11
1 4
Pokazuj tylko zmienne o odpowiedniej skali

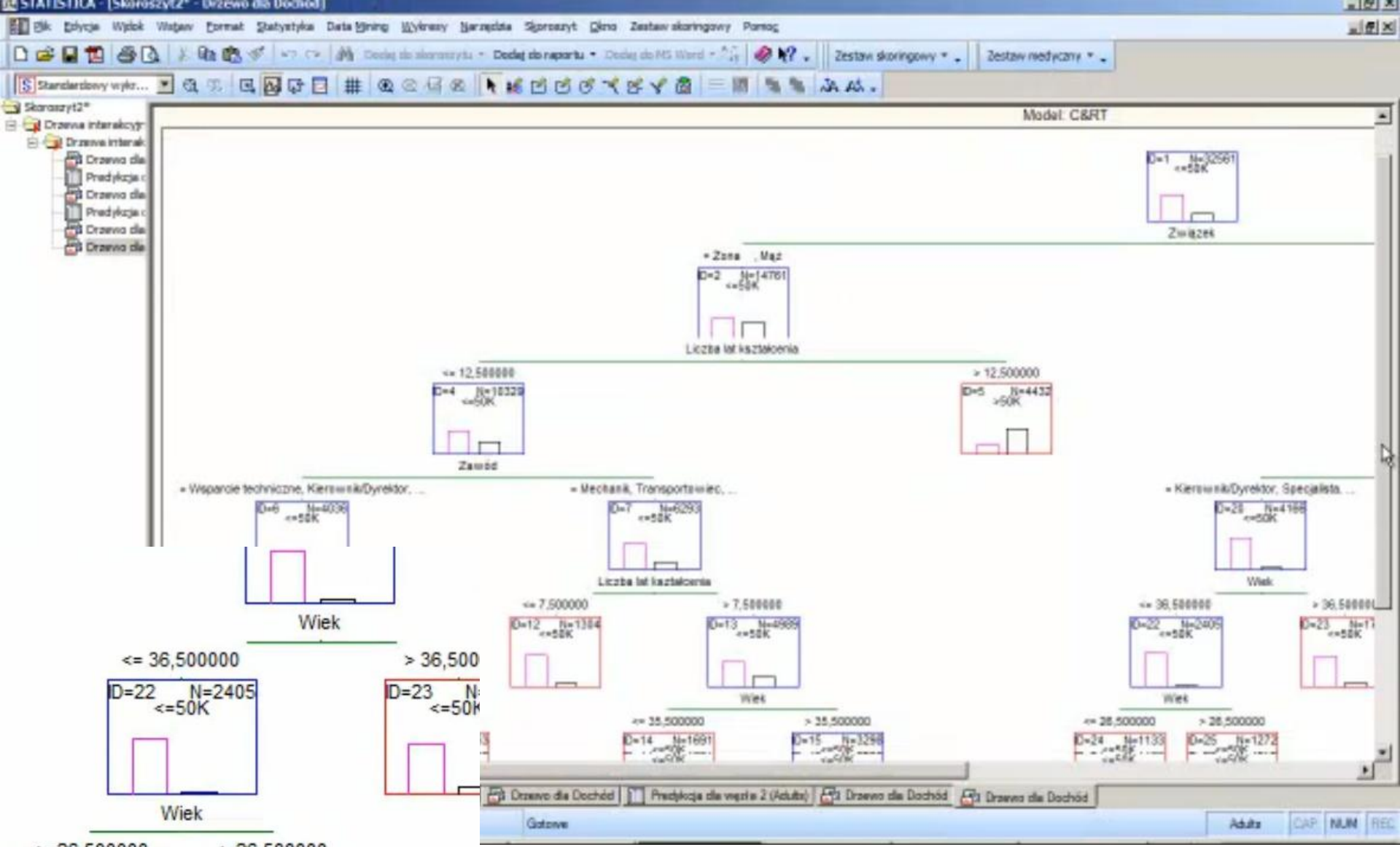

Wyniki drzew interakc. C&RT: Adults
Menedżer
Podstawowe
Klasyfikacja
Predykcja
Raport
Drzewo (budowa, przycinanie):
Buduj drzewo
Eksploruj
Buduj i przytnij drzewo
Usuń wszystkie
Buduj 1 poziom
Usuń 1 poziom
Przegląd drzewa:
Przeglądarka
Przewijalne
Drzewo
Układ drzewa
Węzły i gałęzie:
Węzeł: 1
Określanie 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


<=50K
>50K
Drzewo dla Dochód
Liczba węzłów dzielonych: 0, liczba węzłów końcowych: 1
Model: C&RT
ID=1
N=32561
<=50K

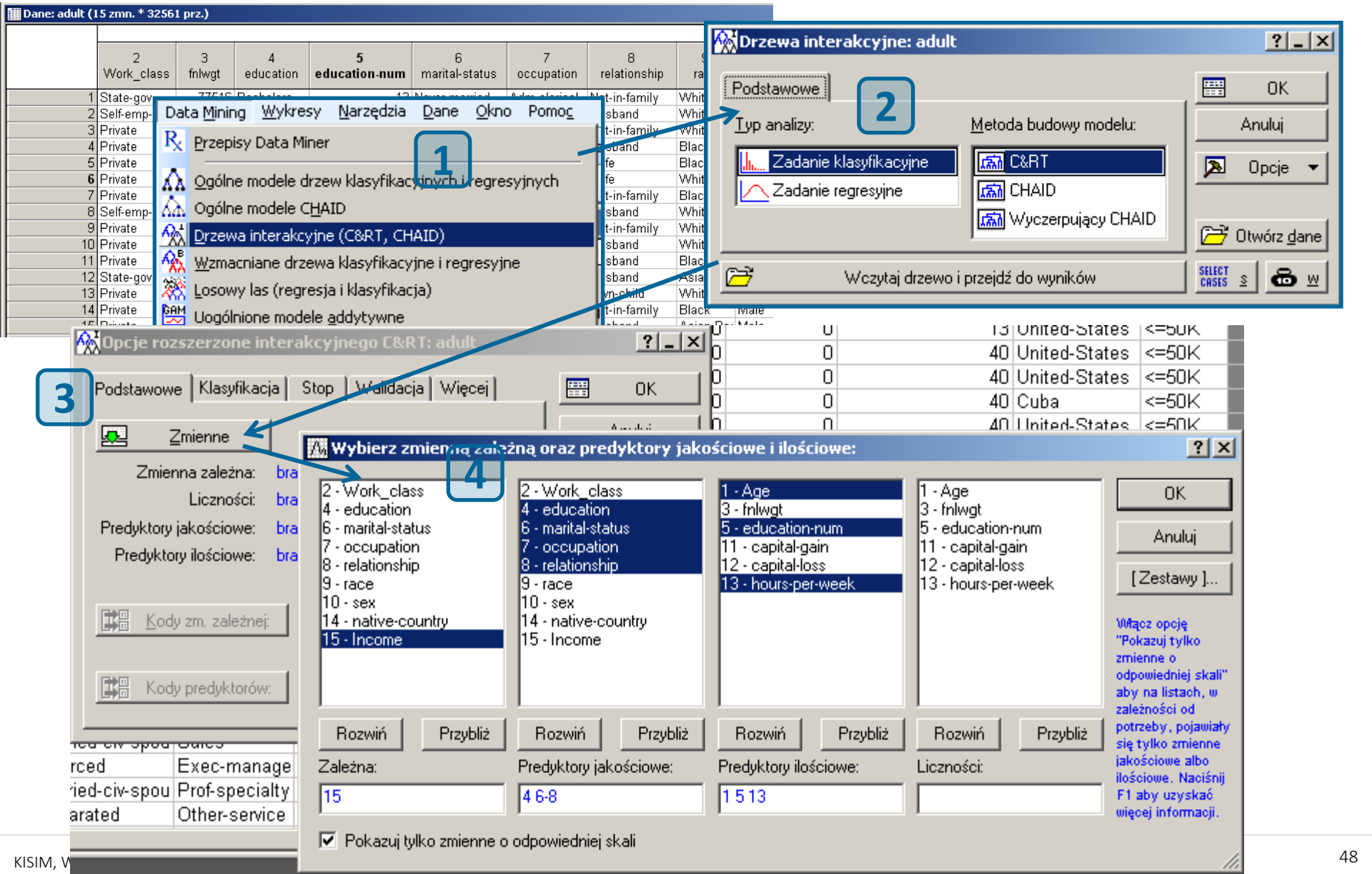

Predykcja dla węzła 1 (Adults)
Model: C&RT
Typ podziału
Poprawa Statystyka
Związek
Automatycznie
0,073548
Stan cywilny
Automatycznie
0,069170
Liczba lat kształcenia
Automatycznie
0,039138
Zawód
Automatycznie
0,034661
Wykształcenie
Wiek
Płeć
Grupa zawodowa
Rasa
Pochodzenia

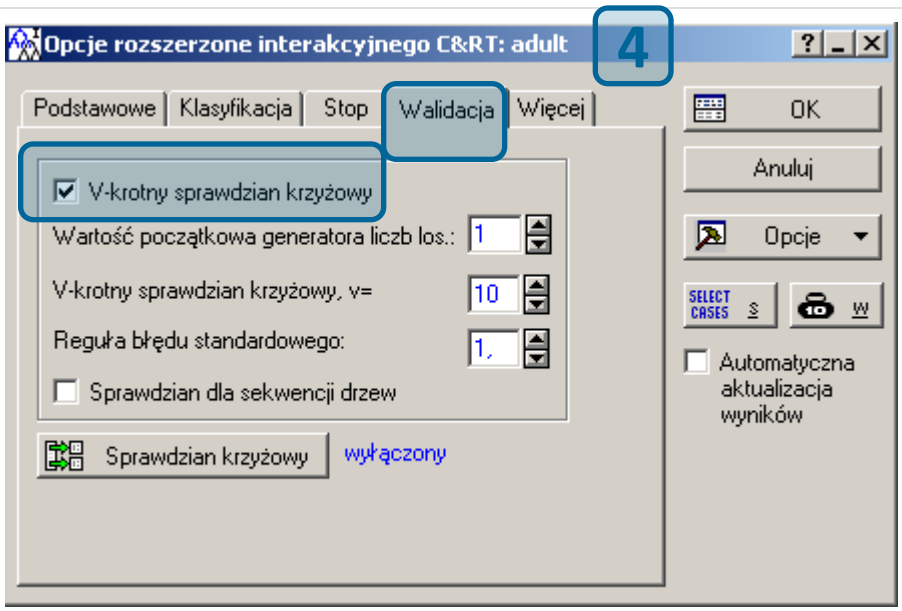

<=50K
>50K
Drzewo dla Dochód
Liczba węzłów dzielonych: 2, liczba węzłów końcowych: 3
Model: C&RT
ID=1
N=32561
<=50K
Związek
= Żona , Mąż
= Poza rodziną, Stanu wolnego, ...
ID=2
N=14761
<=50K
ID=3
N=17800
<=50K
Liczba lat kształcenia
<= 12,500000
> 12,500000
ID=4
N=10329
<=50K
ID=5
N=4432
>50K

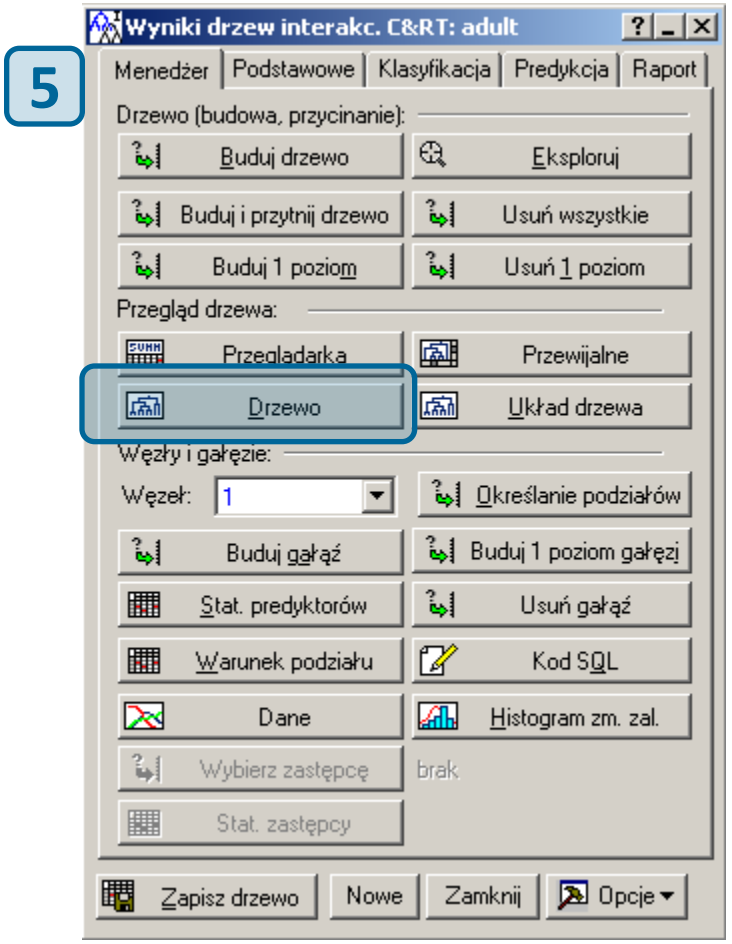

<=50K
>50K
Drzewo dla Dochód
Liczba węzłów dzielonych: 1, liczba węzłów końcowych: 2
Model: C&RT
ID=1
N=32561
<=50K
Związek
= Żona , Mąż
= Poza rodziną, Stanu wolnego, ...
ID=2
N=14761
<=50K
ID=3
N=17800
<=50K

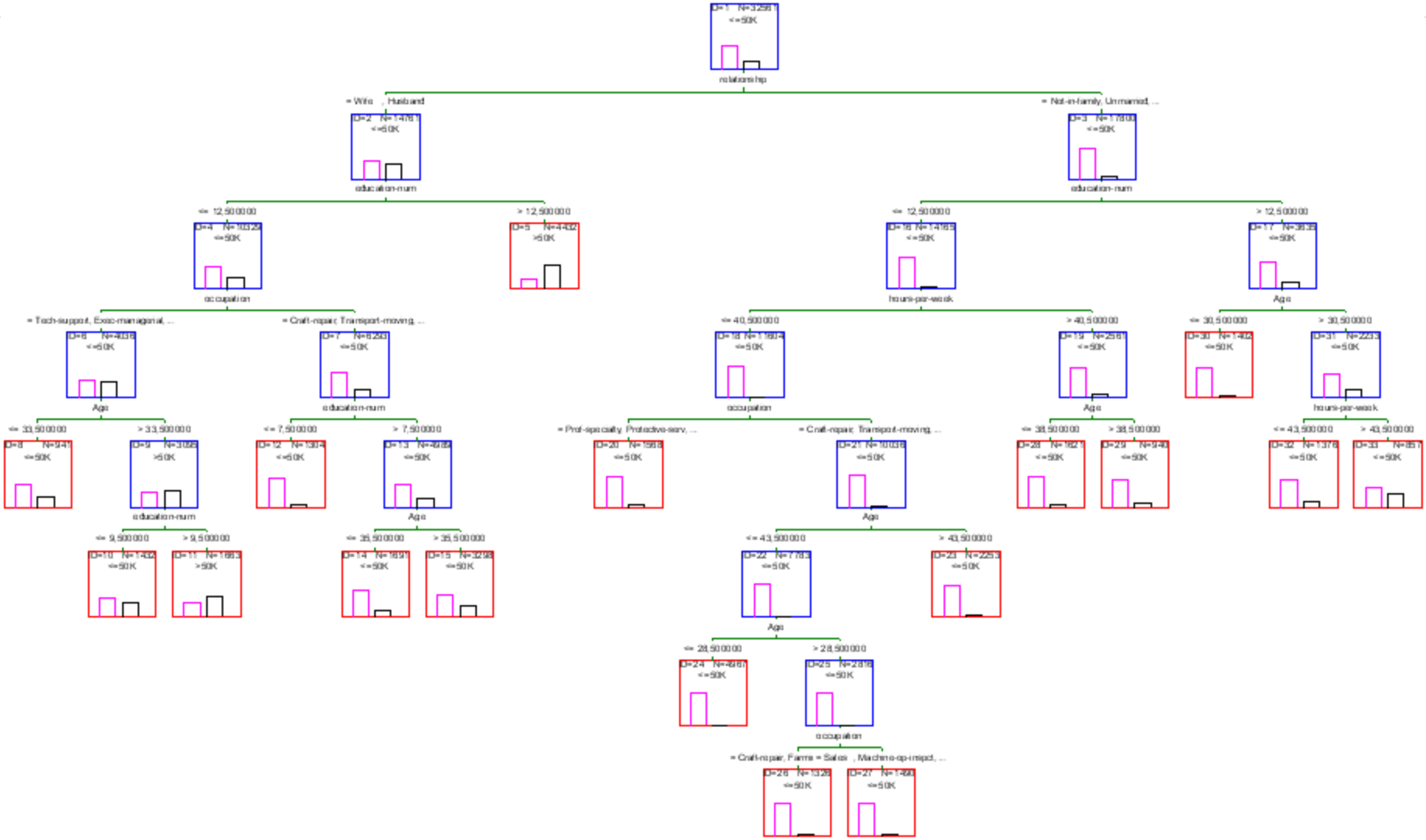

Model: C&RT
Związek
Zona , Mąż
Liczba lat kształcenia
<= 12,500000
> 12,500000
Zawód
Wsparcie techniczne, Kierownik/Dyrektor, ...
Mechanik, Transportowiec, ...
Kierownik/Dyrektor, Specjalista, ...
Liczba lat kształcenia
Wiek
<= 7,500000
> 7,500000
Wiek
Wiek
<= 35,500000
> 35,500000
<= 36,500000
> 36,500
ID=22 N=2405
<=50K
Wiek
<= 26,500000
> 26,500000
ID=24 N=1133
<=50K
ID=25 N=1272
<=50K
Drzewo dla Dochód
Predykcja dla węzła 2 (Adults)
Drzewo dla Dochód
Drzewo dla Dochód
Gotowe
Adults
CAP
NUM
REC

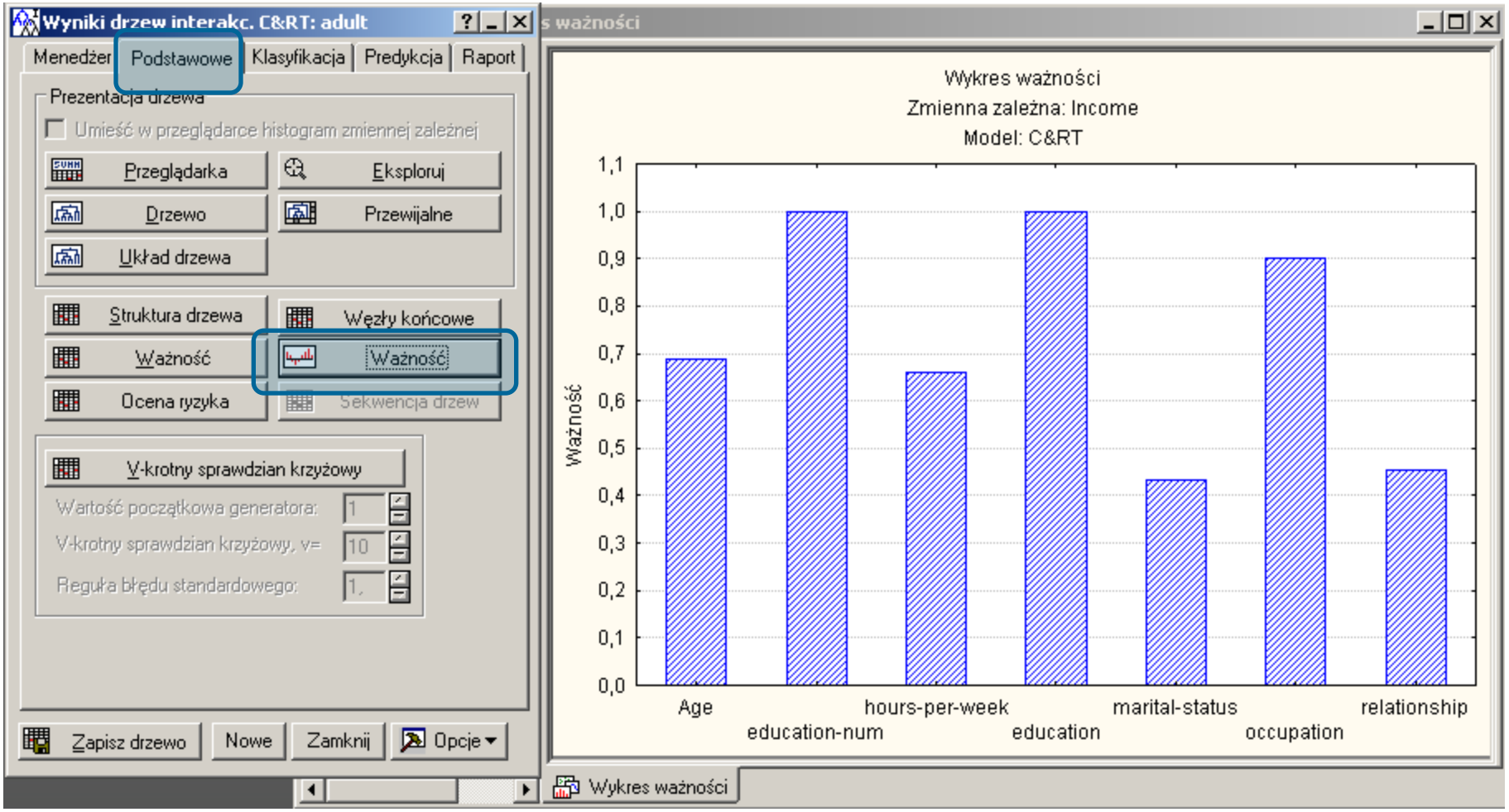
STATISTICA - przykład drzewa C&RT (inny dobór zmiennych)

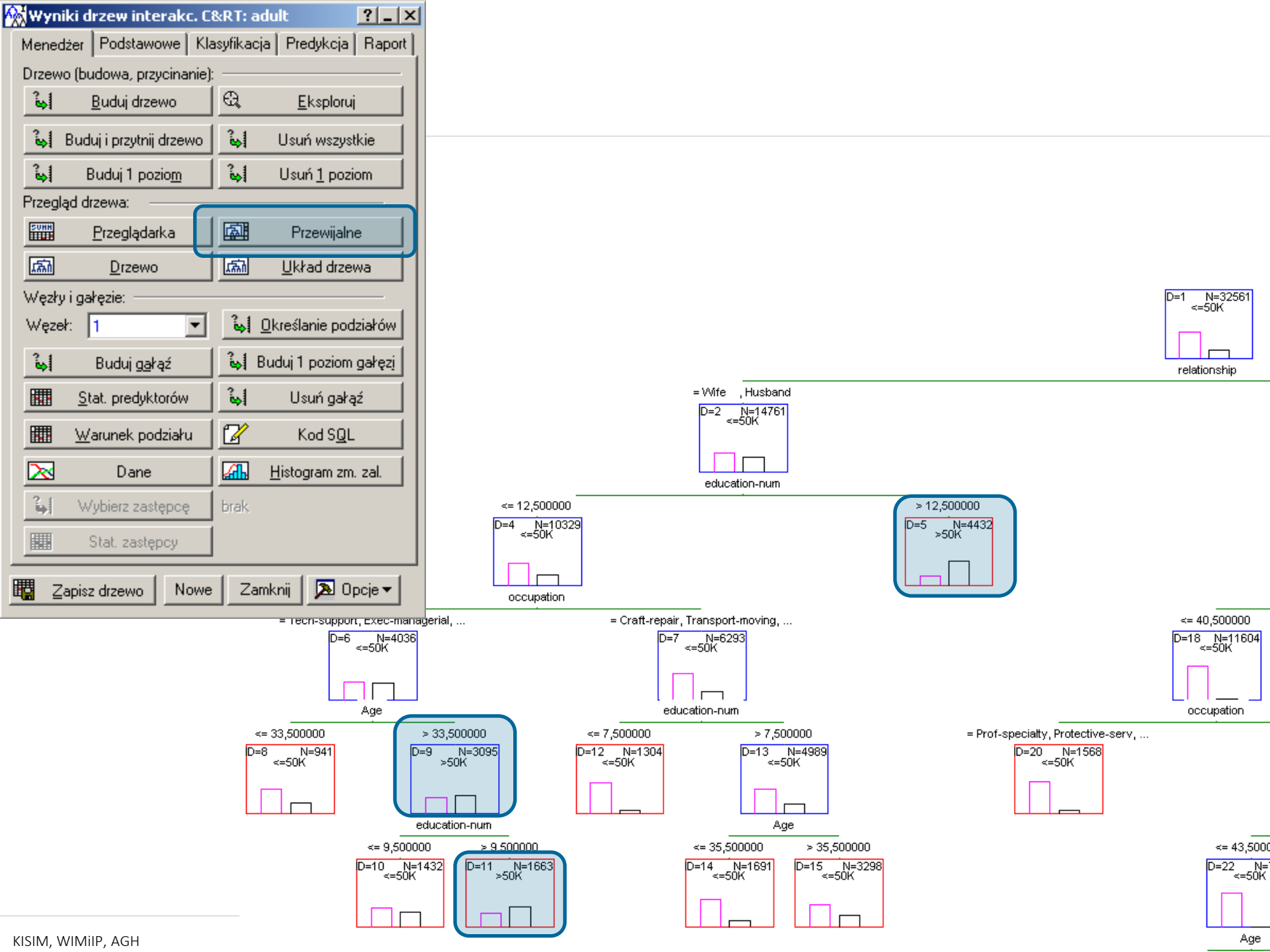


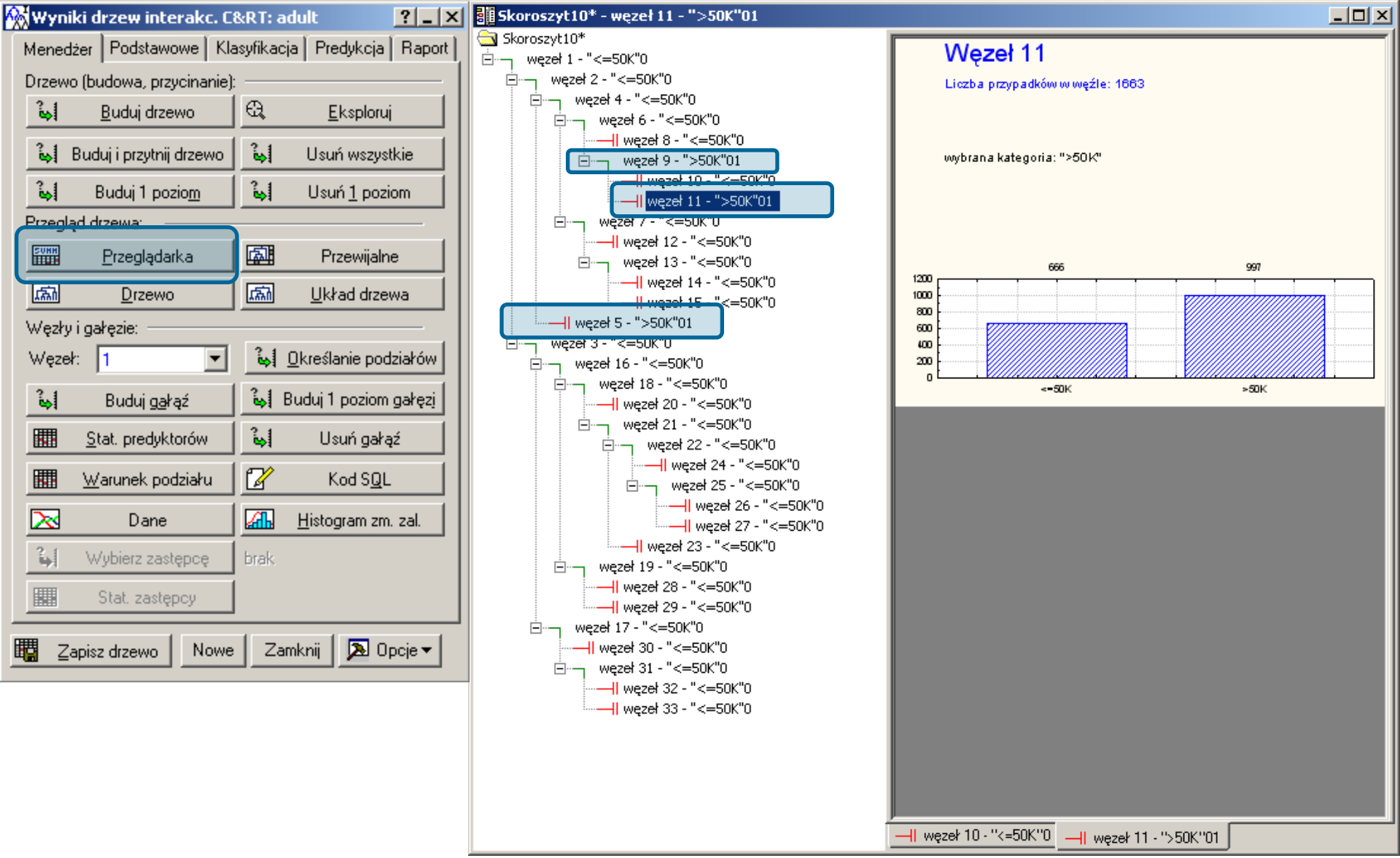

4
Opcje rozszerzone interakcyjnego C&RT: adult
Podstawowe
Klasyfikacja
Stop
Walidacja
Więcej
V-krotny sprawdzian krzyżowy
Wartość początkowa generatora liczb los.:
1
V-krotny sprawdzian krzyżowy, v=
10
Reguła błędu standardowego:
1,
Sprawdzian dla sekwencji drzew
Sprawdzian krzyżowy
wyłączony
OK
Anuluj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


5
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


relationship
= Wife , Husband
= Not-in-family, Unmarried, ...
education-num
education-num
occupation
hours-per-week
Age
= Tech-support, Exec-managerial, ...
= Craft-repair, Transport-moving, ...
Age
education-num
occupation
Age
Age
education-num
education-num
Age
Age
Age
occupation
= Craft-repair, Farms = Sales , Machine-op-inspct, ...

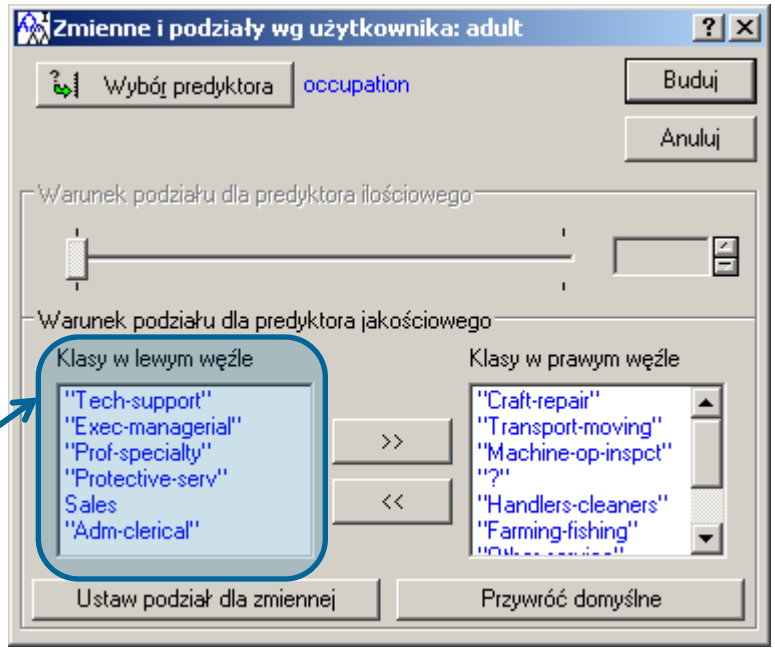

Wyniki drzew interakc. C&RT: adult
Menedżer
Podstawowe
Klasyfikacja
Predykcja
Raport
Prezentacja drzewa
Umieść w przeglądarce histogram zmiennej zależnej
Przeglądarka
Eksploruj
Drzewo
Przewijalne
Układ drzewa
Struktura drzewa
Węzły końcowe
Ważność
Ważność
Ocena ryzyka
Sekwencja drzew
V-krotny sprawdzian krzyżowy
Wartość początkowa generatora:
V-krotny sprawdzian krzyżowy, v=
Reguła błędu standardowego:
Zapisz drzewo
Nowe
Zamknij
Opcje
s 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
<= 43,5000
D=22 N= <=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 10 - "<=50K"0
węzeł 11 - ">50K"01
węzeł 7 - "<=50K"0
węzeł 12 - "<=50K"0
węzeł 13 - "<=50K"0
węzeł 14 - "<=50K"0
węzeł 15 - "<=50K"0
węzeł 5 - ">50K"01
węzeł 3 - "<=50K"0
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

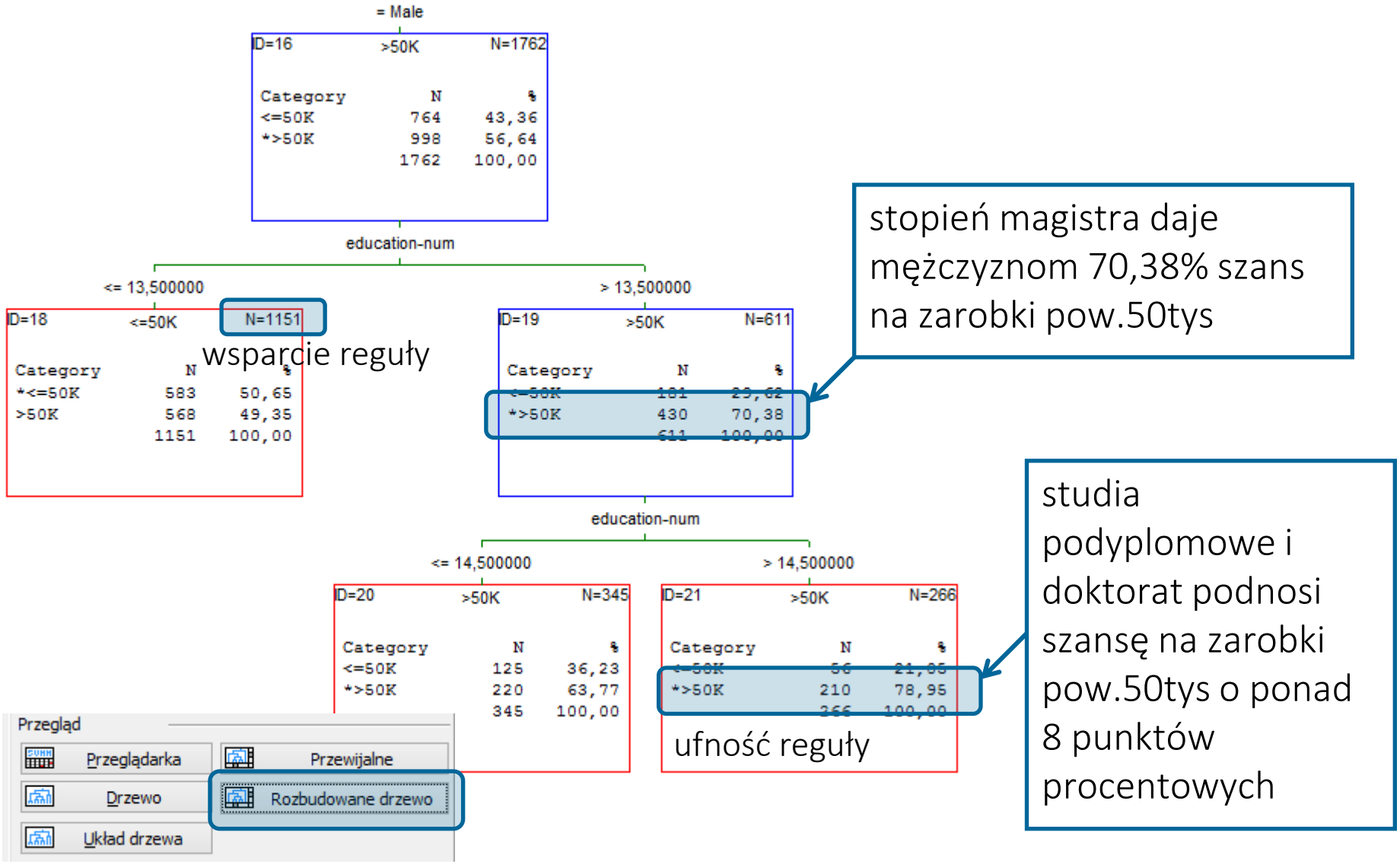
Reguły

- 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%)

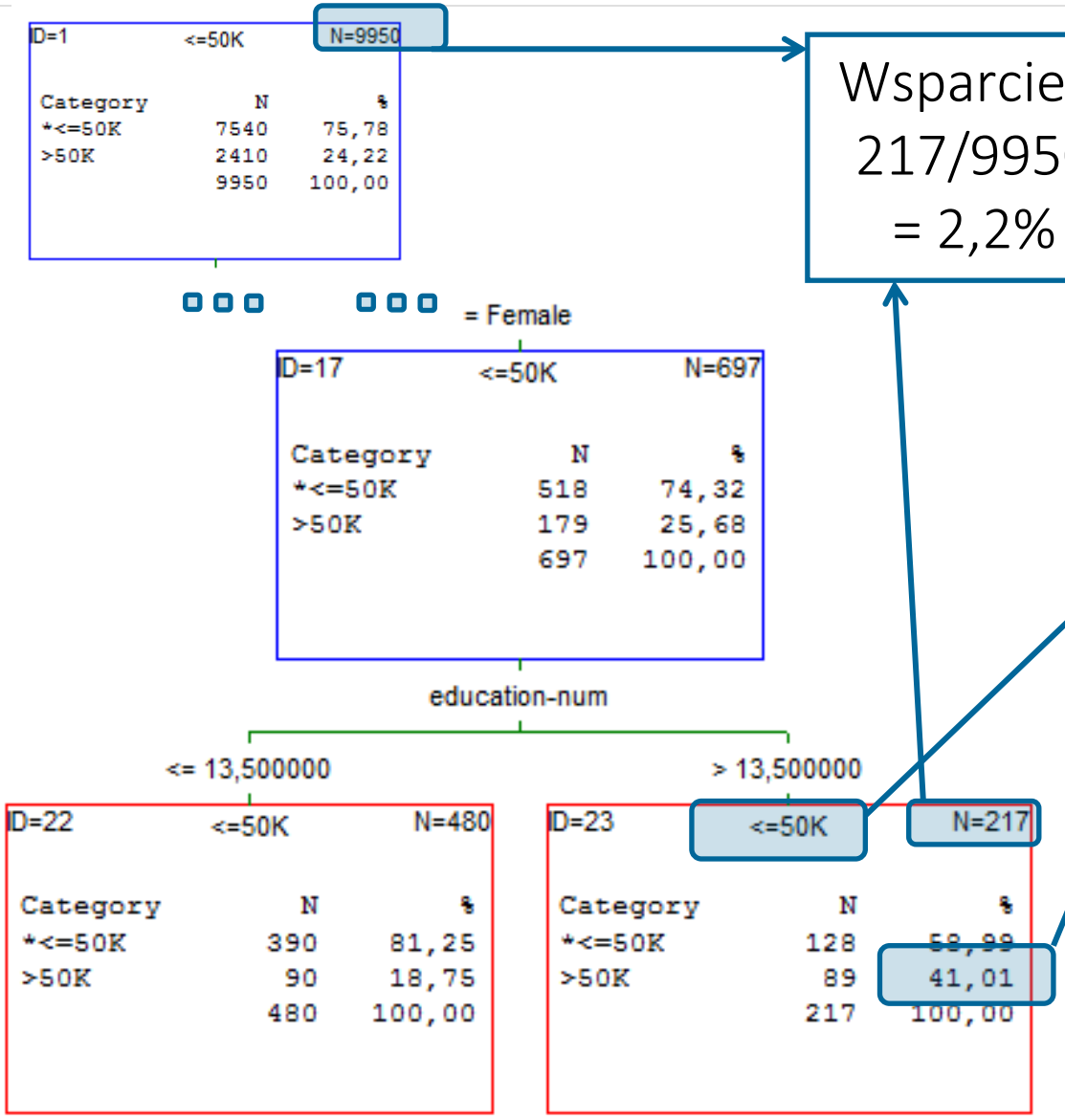


przycisk: „określanie podziałów”

Pewność reguł (ufność reguły, dokładność)



Wsparcie i Ufność



jest bardzo mało kobiet wysoko wykształconych

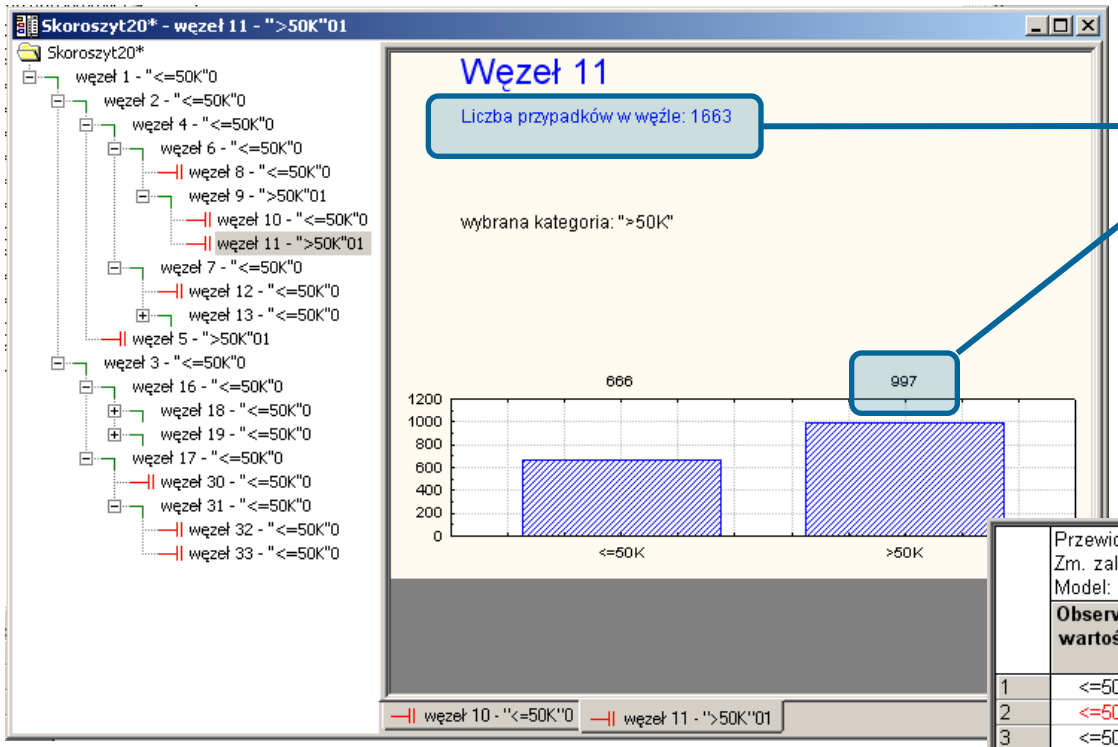
Wsparcie = 217/9950 = 2,2%

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

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

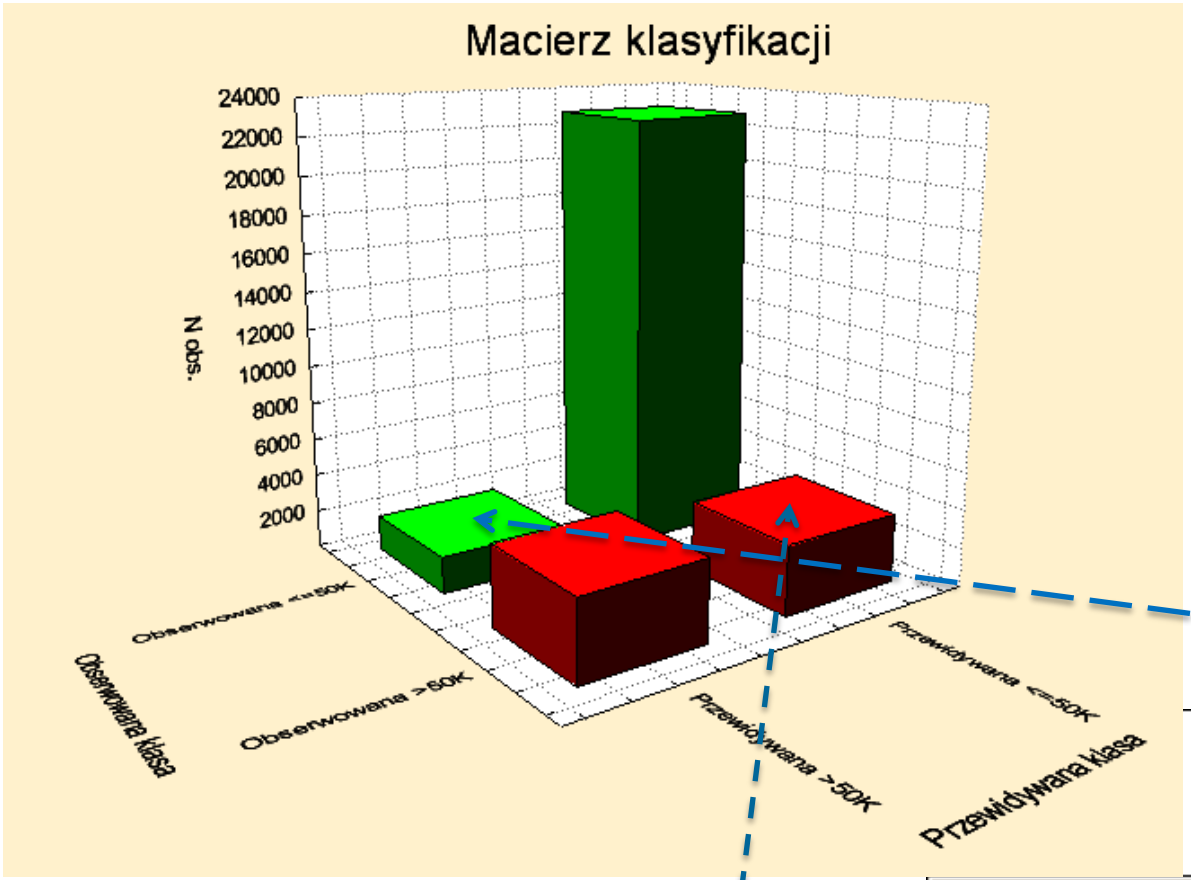
Macierz klasyfikacji (adult) | Macierz klasyfikacji (adult) | Przewidywane (adult) | Wykres przyros

wsparcie reguły:

$N_{węzła} / N_{zbioru}$

ID11: 1633/32561=5%

Macierz klasyfikacji



Ile razy model się pomylił?
Pojęcie „kosztu”

false positives

false negatives

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

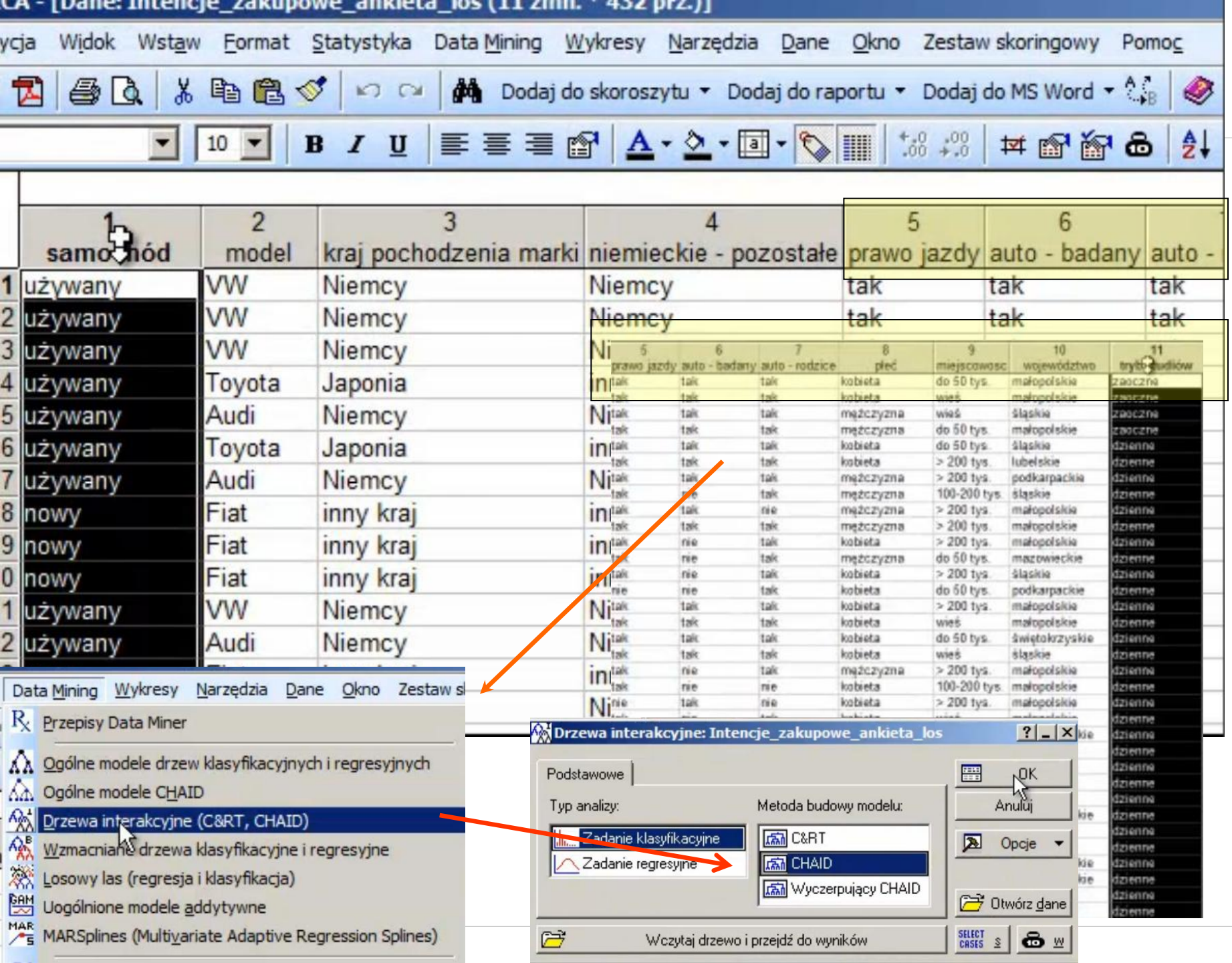
FN

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

Przykład: badanie intencji zakupowych studentów

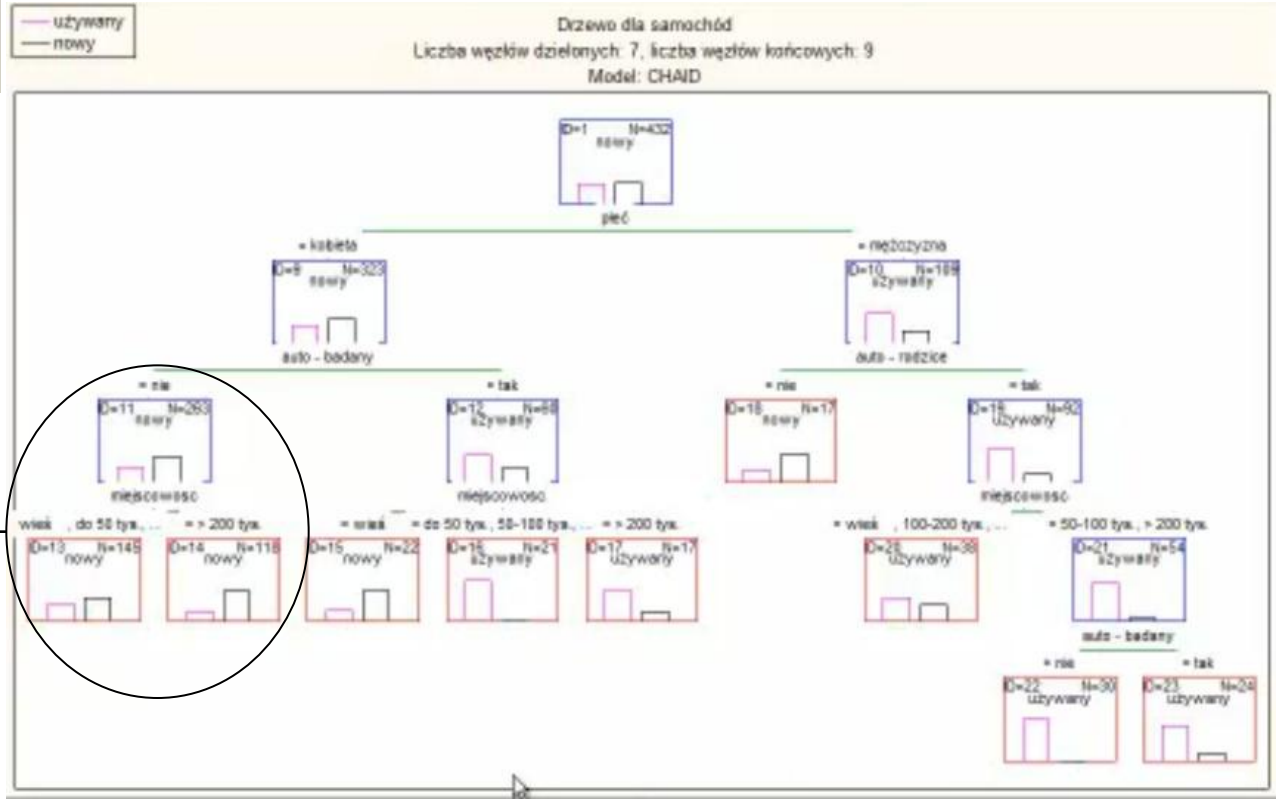
- Co wpływa na skłonność zakupu samochodu nowego bądź używanego?
- Wybór jednego z 12 profili aut o porównywalnej cenie (połowa z nich używane, połowa – nowe)
- 1200 ankietowanych,
- dane demograficzne + wybór

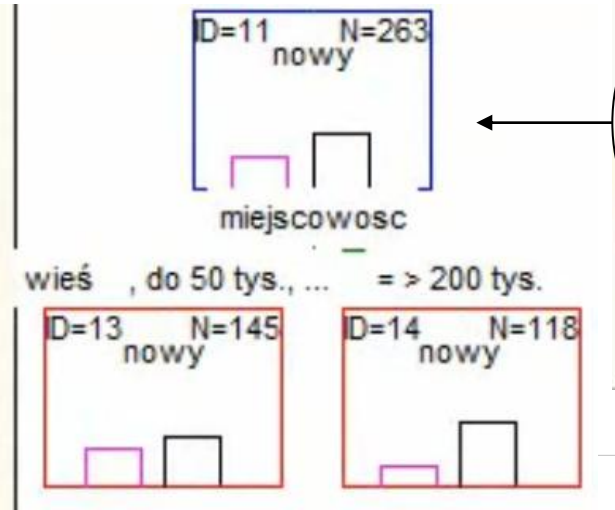

Widok Wstaw Format Statystyka Data Mining Wykresy Narzędzia Dane Okno Zestaw skoringowy Pomoc
Dodaj do skoroszytu
Dodaj do raportu
Dodaj do MS Word
samochód
model
kraj pochodzenia marki
niemieckie - pozostałe
prawo jazdy
auto - badany
auto - rodzice
płeć
miejscowość
województwo
tryb budków
Data Mining
Wykresy
Narzędzia
Dane
Okno
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

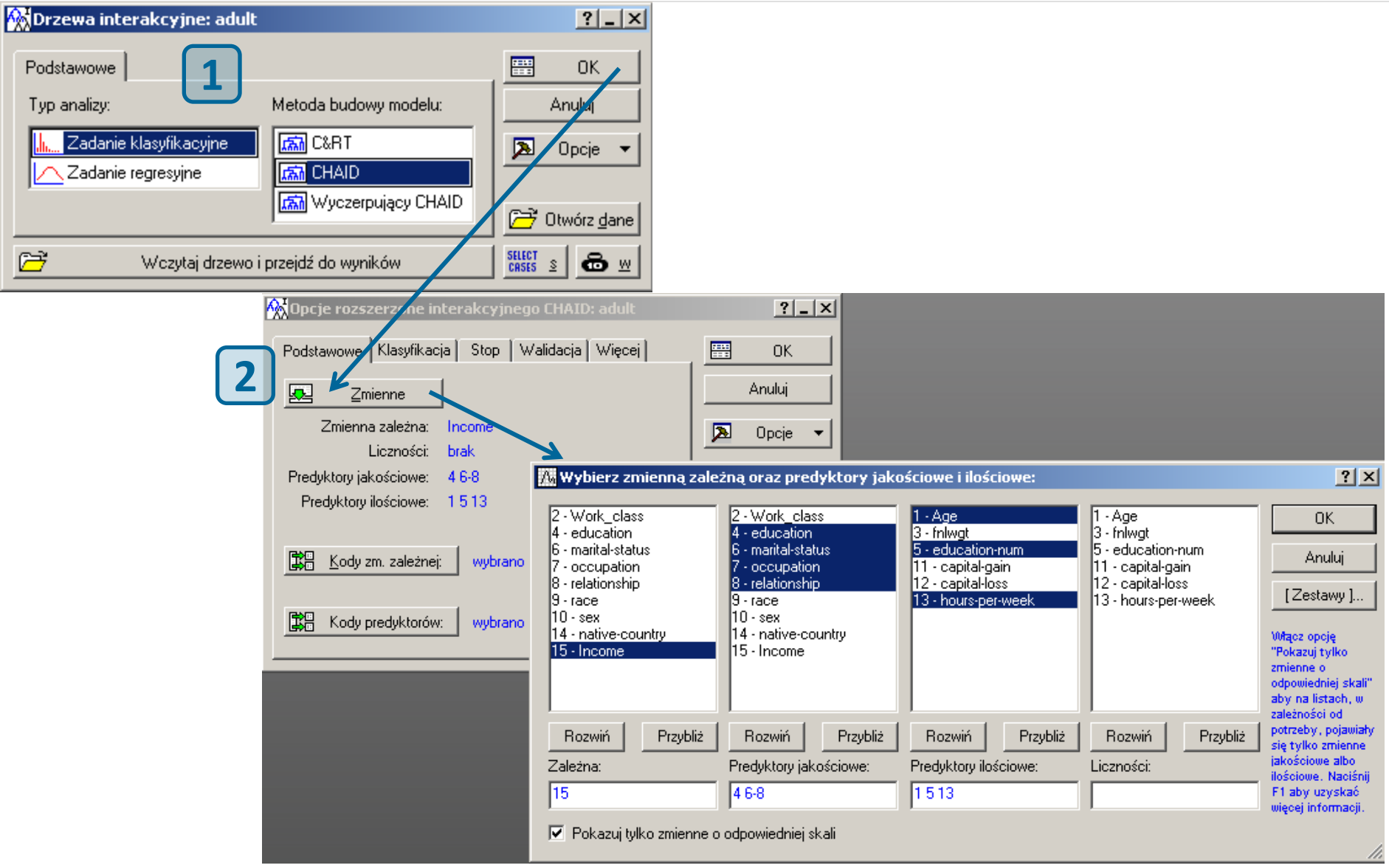
	Liczba węzły	Typ podziału	chi-kwadrat Statystyka	Stopnie Swobody	Skorygowane p
płeć	2	Automatycznie	27,13573	1,000000	0,000000
prawo jazdy	2	Automatycznie	21,86719	1,000000	0,000003
auto - badany	2	Automatycznie	17,78424	1,000000	0,000025
tryb studiów	3	Automatycznie	28,19740	2,000000	0,000001
województwo	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

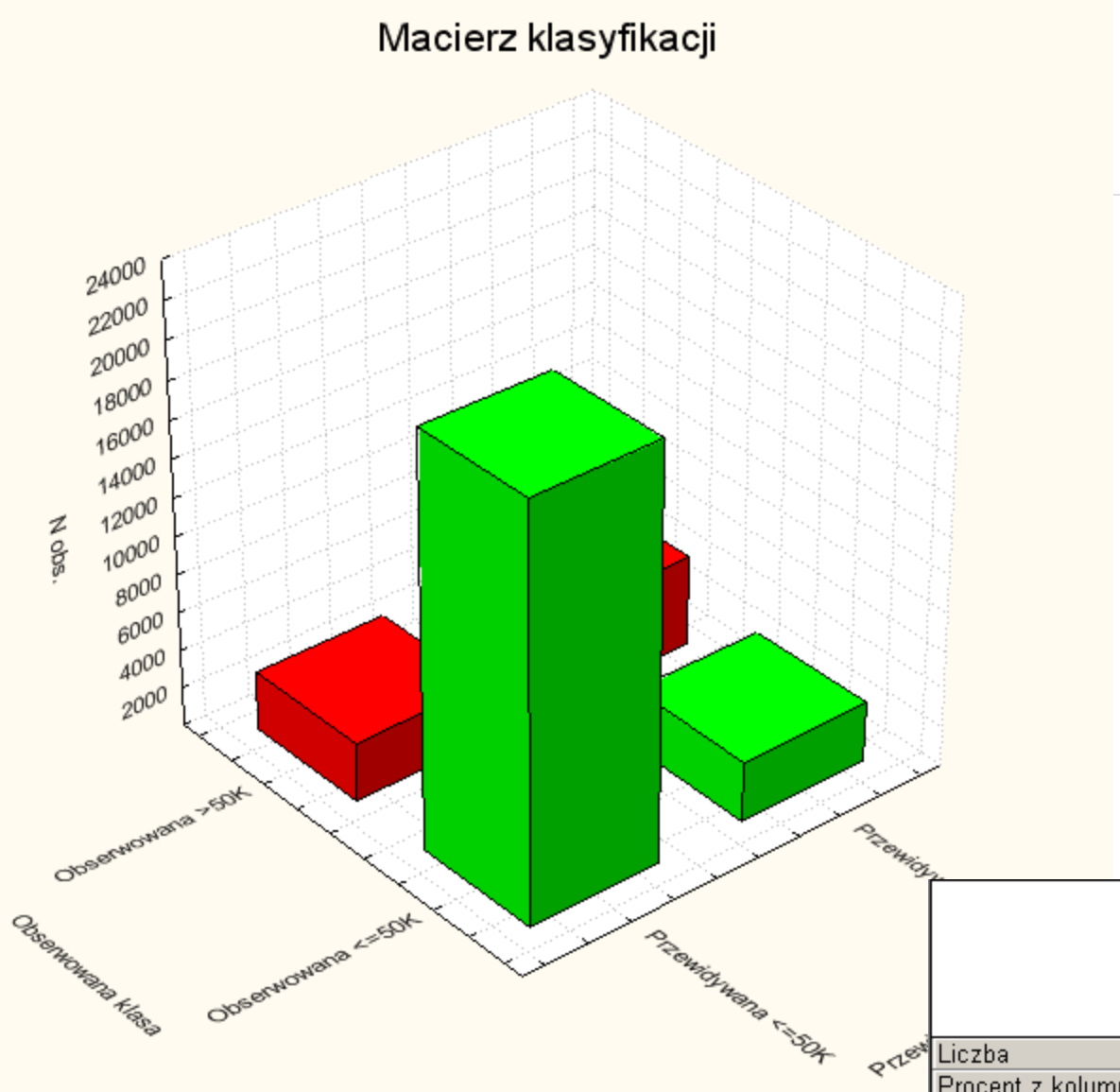




STATISTICA - inny przykład drzewa CHAID



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

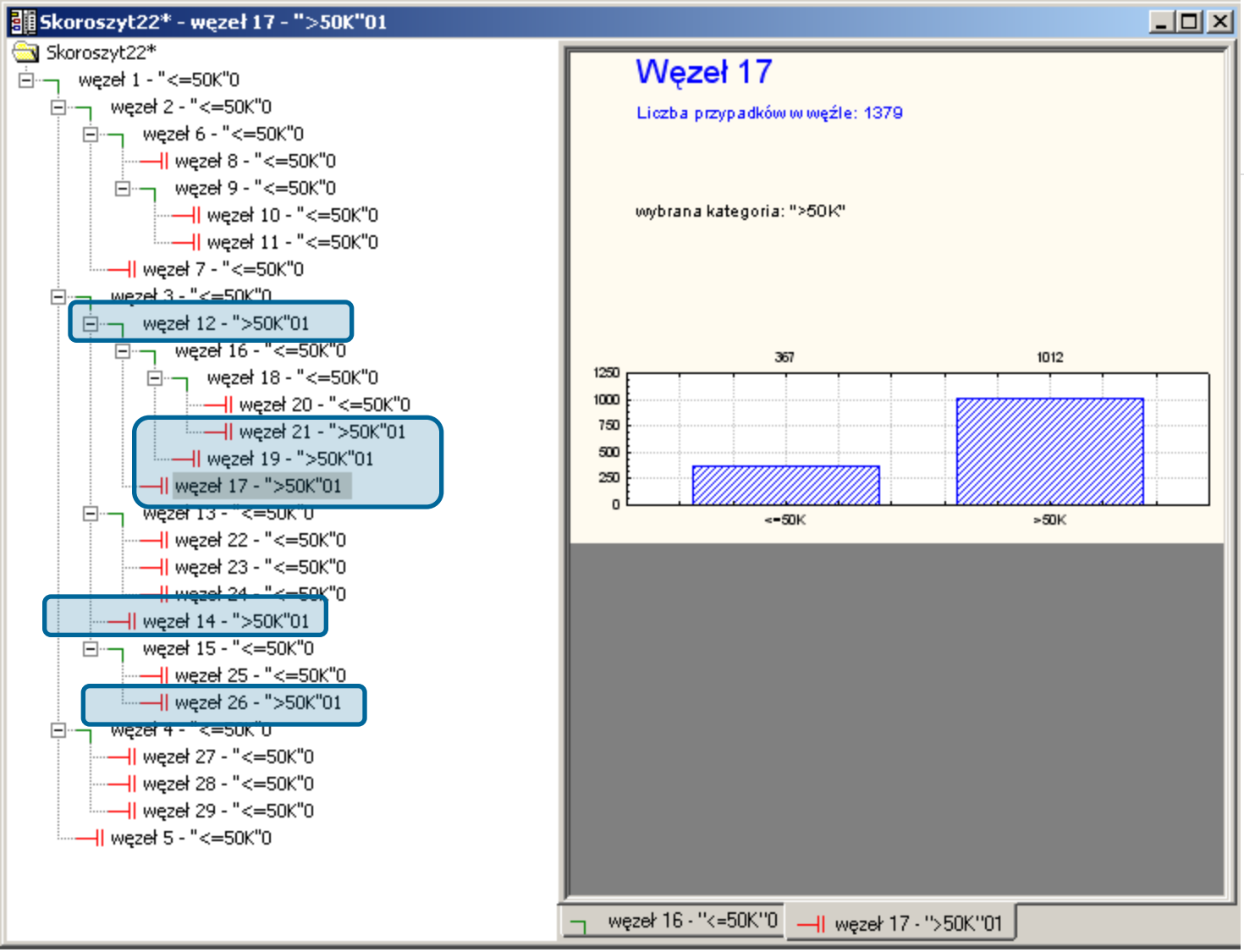


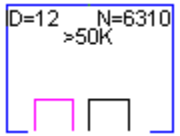
Macierz klasyfikacji (adult)
Zm. zal.:Income
Model: CHAID

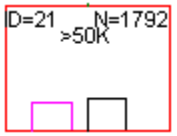
	Obserw.	Przewidywana <=50K	**Przewidywana >50K**	Łącznie w wierszu
Liczba	<=50K	21569	3151	24720
Procent z kolumnny		87.63%	39.65%	
Procent z wiersza		87.25%	12.75%	
Procent z ogółu		66.24%	9.68%	75.92%
Liczba	>50K	3045	4796	7841
Procent z kolumnny		12.37%	60.35%	
Procent z wiersza		38.83%	61.17%	
Procent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

FP 2

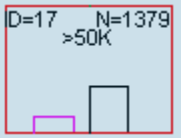
FN 1

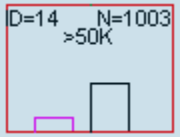

Skoroszyt22* - węzeł 17 - ">50K"01
Skoroszyt22*
węzeł 1 - "<=50K"0
węzeł 2 - "<=50K"0
węzeł 6 - "<=50K"0
węzeł 8 - "<=50K"0
węzeł 9 - "<=50K"0
węzeł 10 - "<=50K"0
węzeł 11 - "<=50K"0
węzeł 7 - "<=50K"0
węzeł 3 - "<=50K"0
węzeł 12 - ">50K"01
węzeł 16 - "<=50K"0
węzeł 18 - "<=50K"0
węzeł 20 - "<=50K"0
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"
367
1012
1250
1000
750
500
250
0
<=50K
>50K
węzeł 16 - "<=50K"0
węzeł 17 - ">50K"01

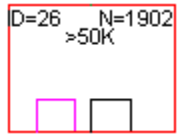

D=12 N=6310
>50K

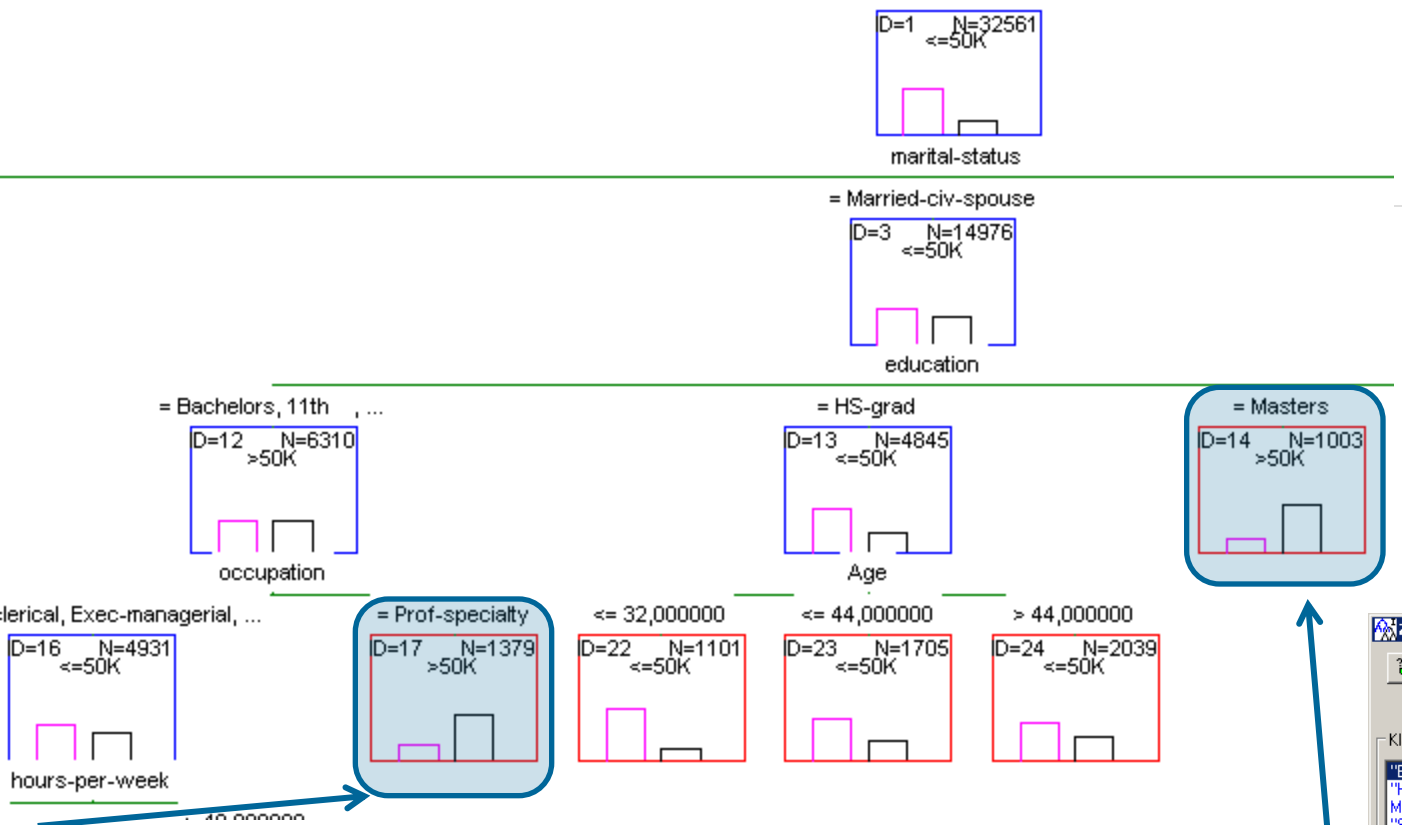

D=21 N=1792
>50K

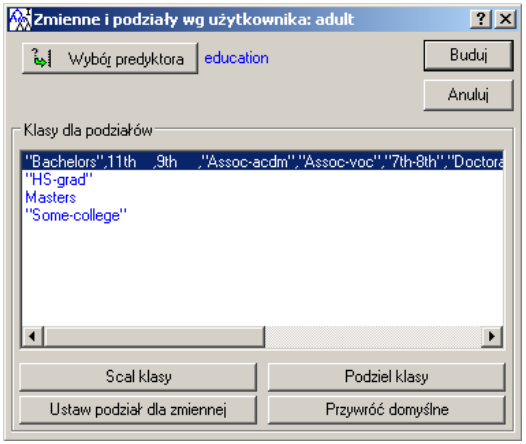

D=19 N=1871
>50K


D=17 N=1379
>50K


D=14 N=1003
>50K


D=26 N=1902
>50K





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%)

zalety drzew

- łatwe w interpretacji, interpretacja w postaci reguł
- wyjaśniają zależności (*white-box*)
- pozwalają klasyfikować nieliniowe problemy
- szybko się uczą
- zmienne ilościowe, jak i jakościowe
- nie wymagają założeń co do rozkładu
- pozwalają określić ważność predyktorów, niewrażliwe na nieistotne zmienne i wzajemne korelacje
- pełne pokrycie przestrzeni wyników
- niewrażliwość na wartości odstające – podział w punkcie, nawet jeśli jakieś zmienne osiągają bardzo wysokie/niskie wartości
- radzenie sobie z brakami danych – podziały zastępcze

wady drzew

- duża wariancja (mała zmiana w zbiorze danych może wywołać dużą zmianę w tworzonych podziałach)
- skłonność do przeuczania (bez przycinania drzewa uczą się złożonych, przypadkowych wzorców)
- wymagają wyczucia w ustalaniu kosztu dla FN i FP
- *bias* w przypadku nierównej częstości klas
- brak zdolności do ekstrapolacji (regresyjne)
- podejście zachłanne, brak mechanizmów szukania optimów globalnych w szukaniu podziałów

ensemble methods

- uczenie zespołowe drzew: Multiple classifiers

Homogeneous classifiers – ten sam algorytm, wiele modeli

» Bagging (Breiman): Bootstrap aggregation

» Boosting (Freund, Schapire): AdaBoost, changing the distribution of training examples

» Multiple partitioned data

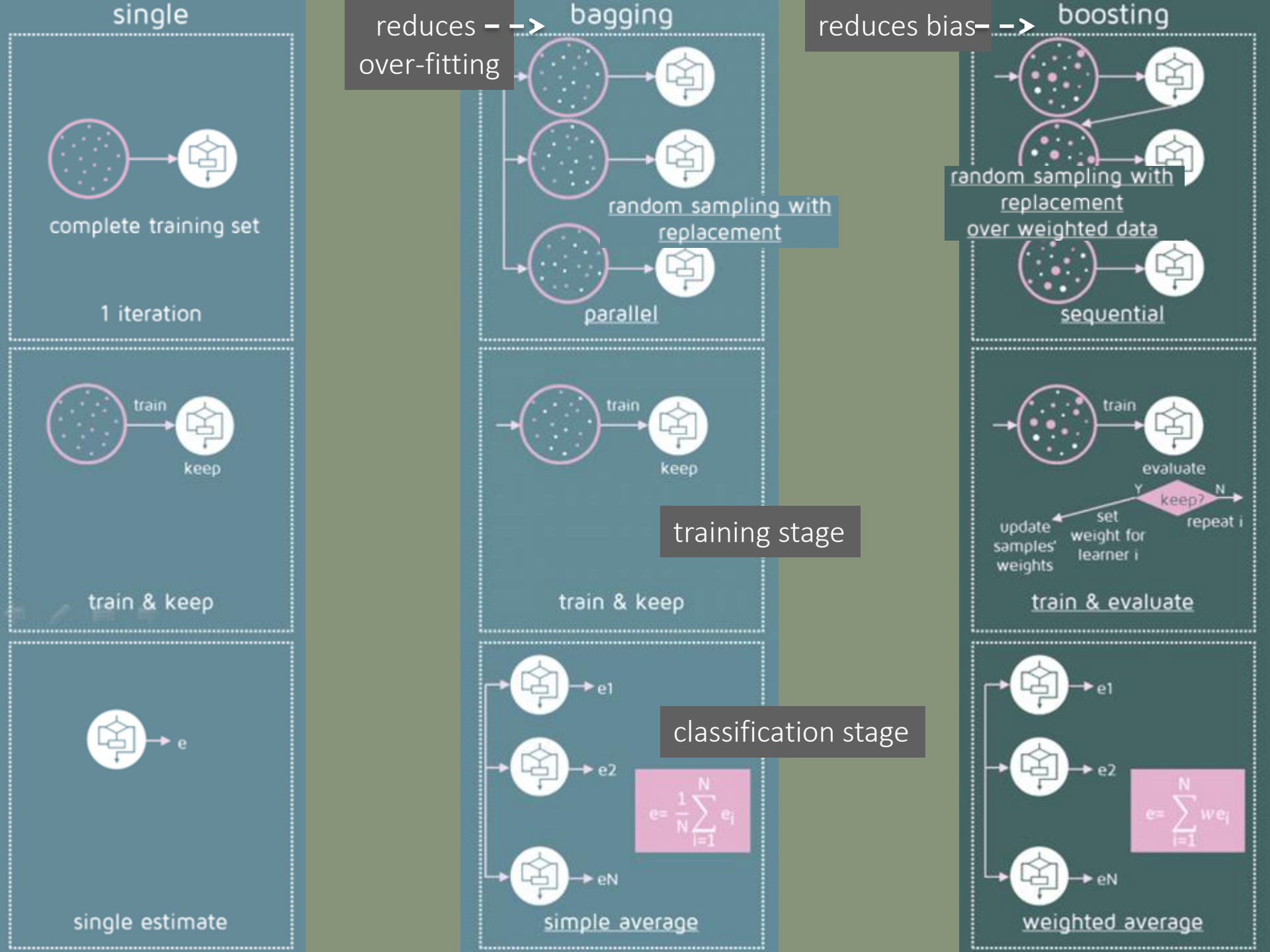
» Multi-class specialized systems, (e.g. ECOC pairwise classification)

Bagging (Bootstrap Aggregation)

- celem jest zmniejszenie wariancji drzew
- budowane jest kilka modeli na różnych podzbiorach (losowanie bootstrapowe)
- niezależne drzewa dokonują predykcji, a wynik jest uśredniany
- przykład: Random Forest (losowy las)
- w przypadku losowego lasu losowane są nie tylko przypadki, ale również zmienne

Boosting (drzewa wzmacniane)

- również tworzy zbiór modeli
- podejście sekwencyjne: tworzone są proste drzewa, każde kolejne ma na celu zmniejszać błąd poprzednich
- w każdym kroku rosną wagi przypadków, które zostały błędnie sklasyfikowane
- w ten sposób najtrudniejsze przypadki uzyskują największe znaczenie, a kolejne drzewa zwiększają precyzję
- przykłady wzmacnianych drzew: Gradient Boost, XGBoost, AdaBoost, etc.


single
reduces over-fitting
bagging
reduces bias
boosting
complete training set
1 iteration
random sampling with replacement
parallel
random sampling with replacement over weighted data
sequential
train
keep
train & keep
train
keep
train & keep
train
evaluate
keep?
Y
N
repeat i
update samples' weights
set weight for learner i
train & evaluate
training stage
e
single estimate
e1
e2
eN
$e = \frac{1}{N}\sum_{i=1}^{N} e_i$
simple average
classification stage
e1
e2
eN
$e = \sum_{i=1}^{N} we_i$
weighted average