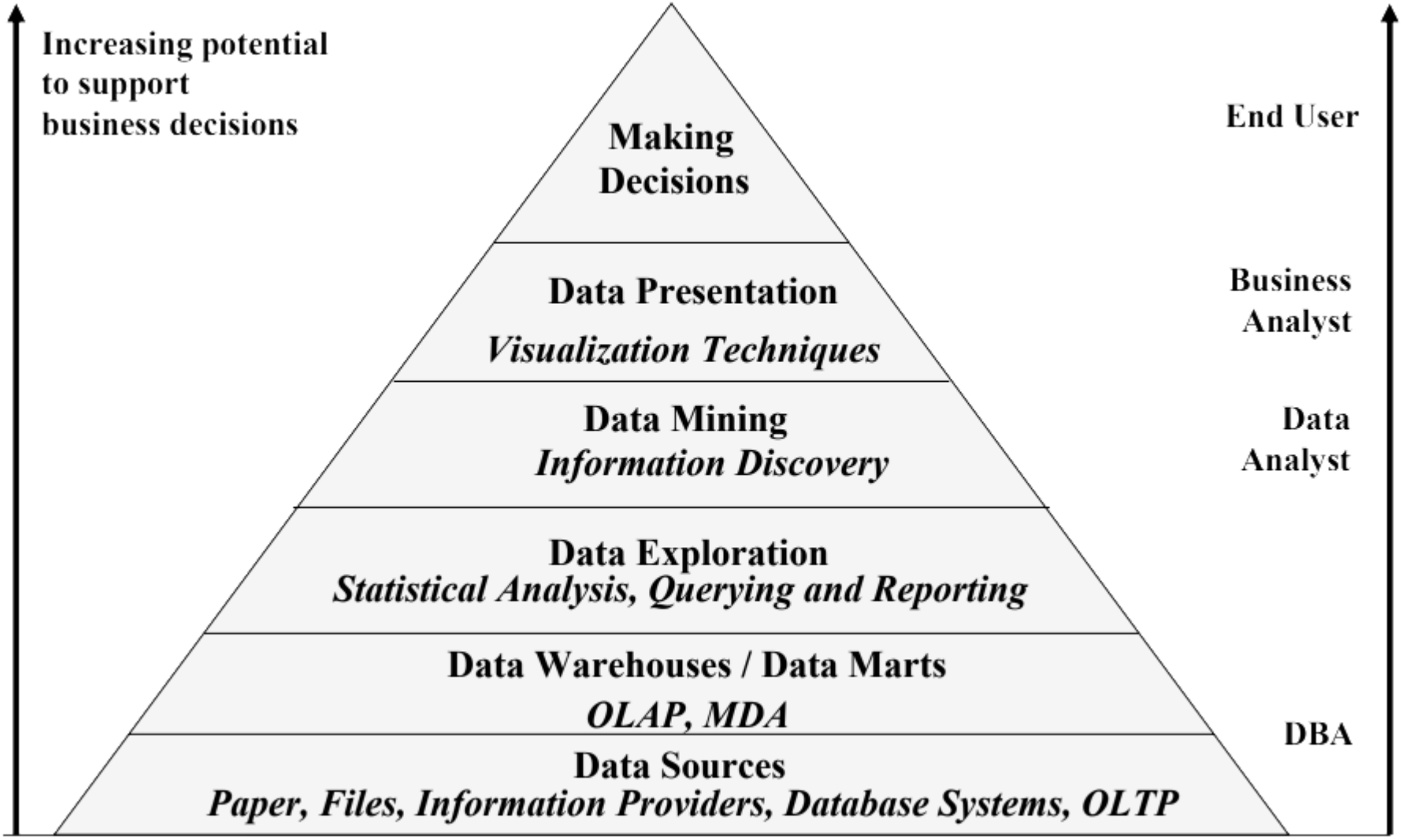
Indukcja reguł

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

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

Data Mining and Business Intelligence



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…

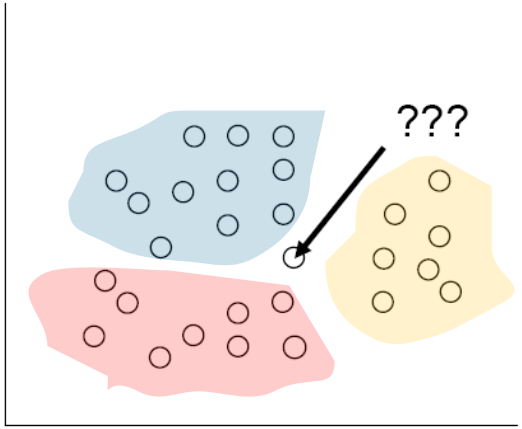
- proste algorytmy często wystarczają
- jest wiele przypadków prostych struktur danych:
  - jeden atrybut załatwia sprawę
  - wszystkie atrybuty mają podobny wpływ i są niezależne
  - logiczna struktura i mała liczba atrybutów odpowiednich dla drzew
  - zbiór kilku reguł logicznych
  - zależności pomiędzy 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)

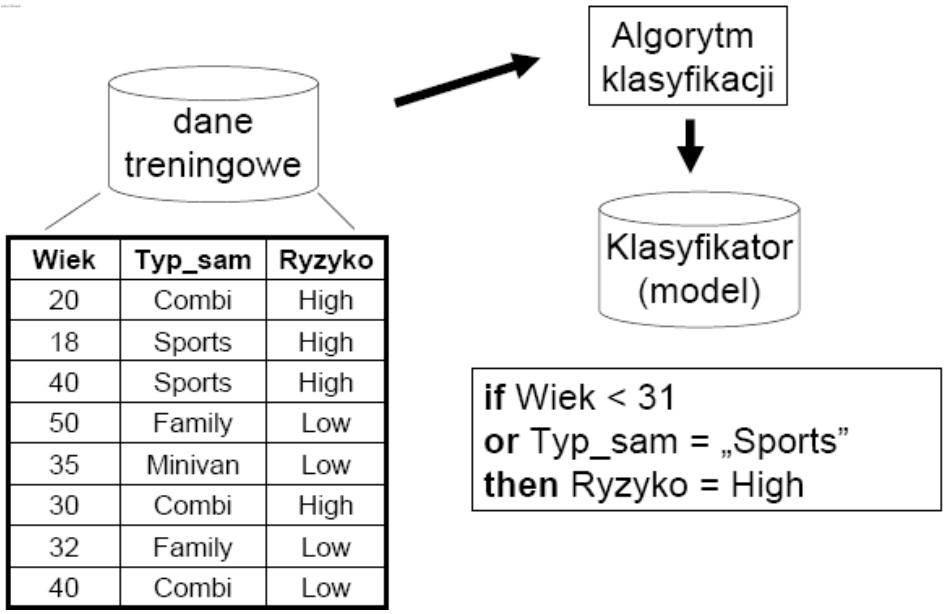
– **Dane wejściowe**

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

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

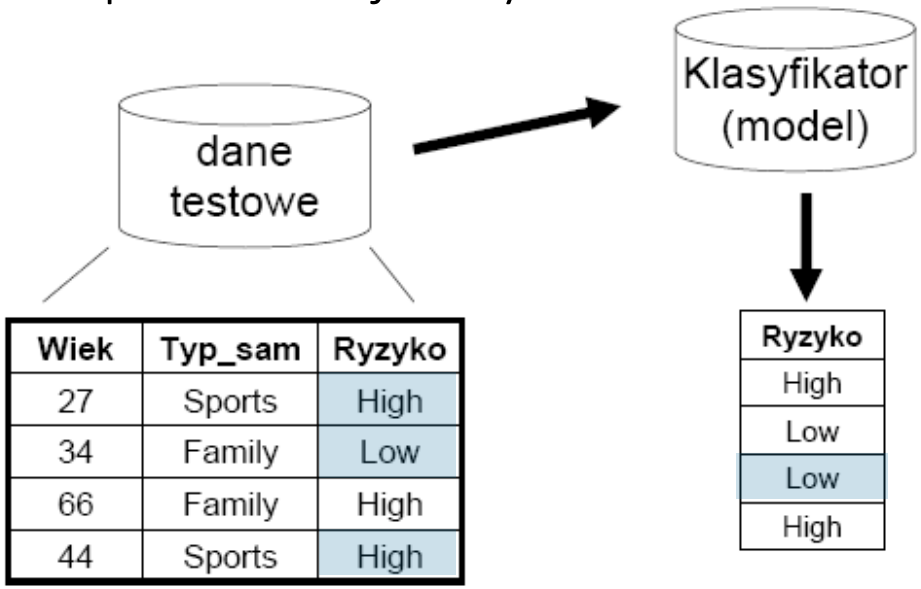
- Wynik klasyfikacji:
  - » Reguły klasyfikacyjne postaci *IF - THEN*
  - » Drzewa decyzyjne
- Istotną sprawą z punktu widzenia poprawności i efektywności modelu jest tzw. dokładność modelu.
- dla przykładów testowych, dla których znane są wartości atrybutu decyzyjnego, wartości te są porównywane z wartościami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

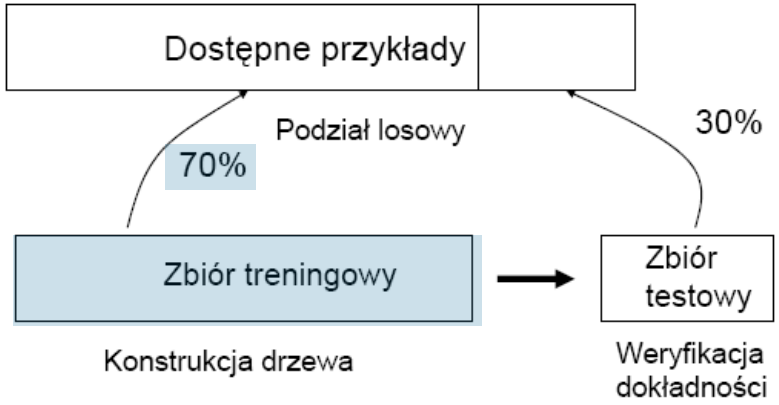
Jeżeli dokładność klasyfikatora jest akceptowalna, wówczas możemy wykorzystać klasyfikator do klasyfikacji nowych danych.

Celem klasyfikacji, jak pamiętamy jest przyporządkowanie nowych danych dla których wartość atrybutu decyzyjnego nie jest znana do odpowiedniej klasy.



Testowanie

Duży zbiór danych



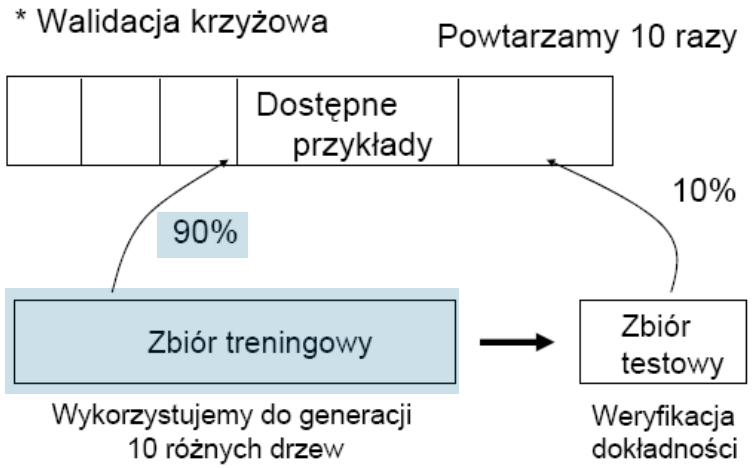
Mały zbiór danych

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)

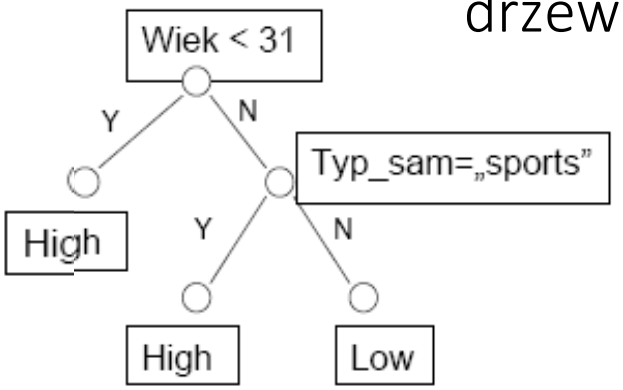
Rodzaje modeli klasyfikacyjnych:

- 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

Klasyfikacja poprzez indukcję drzew decyzyjnych



Drzewo decyzyjne jest grafem o strukturze drzewiastej, gdzie

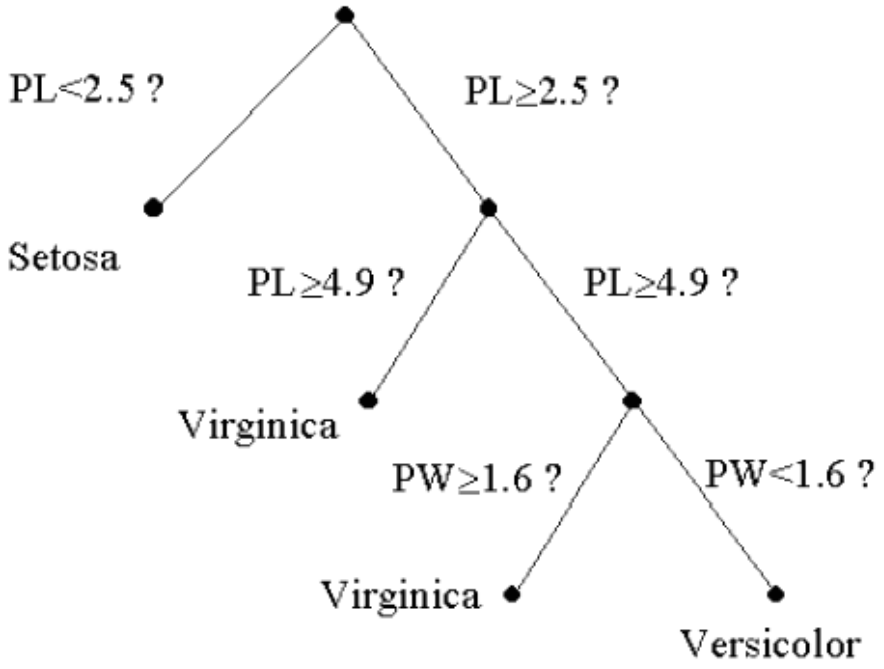
- każdy wierzchołek wewnętrzny reprezentuje test na atrybucie (atrybutach),
- każdy łuk reprezentuje wynik testu,
- każdy liść reprezentuje pojedynczą klasę lub rozkład wartości klas

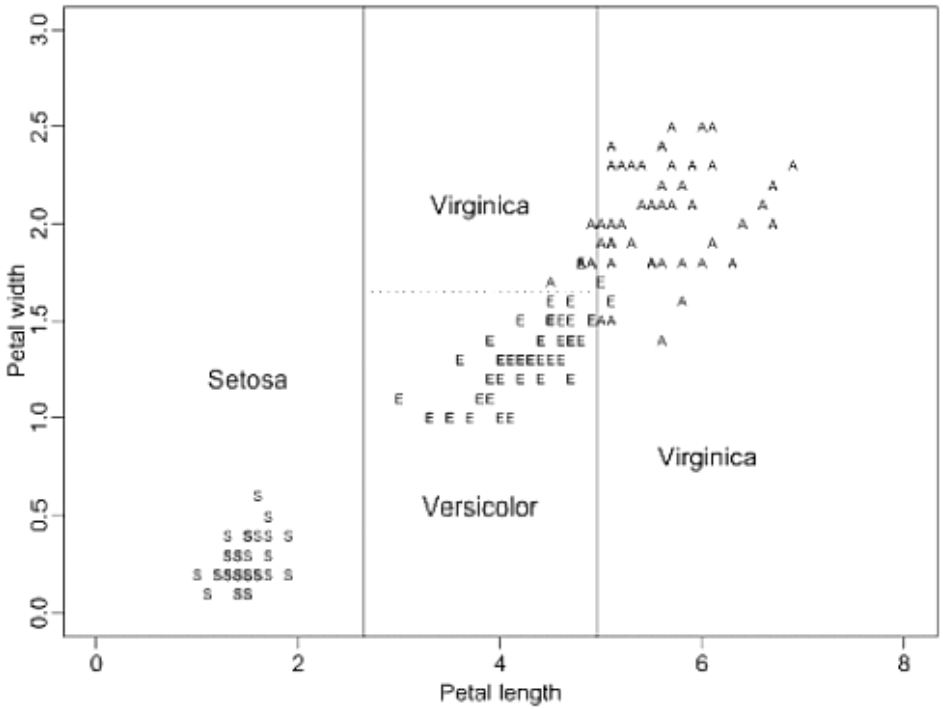
Drzewo decyzyjne dzieli zbiór treningowy na partycje do momentu, w którym każda partycja zawiera dane należące do jednej klasy, lub, gdy w ramach partycji dominują dane należące do jednej klasy

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

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

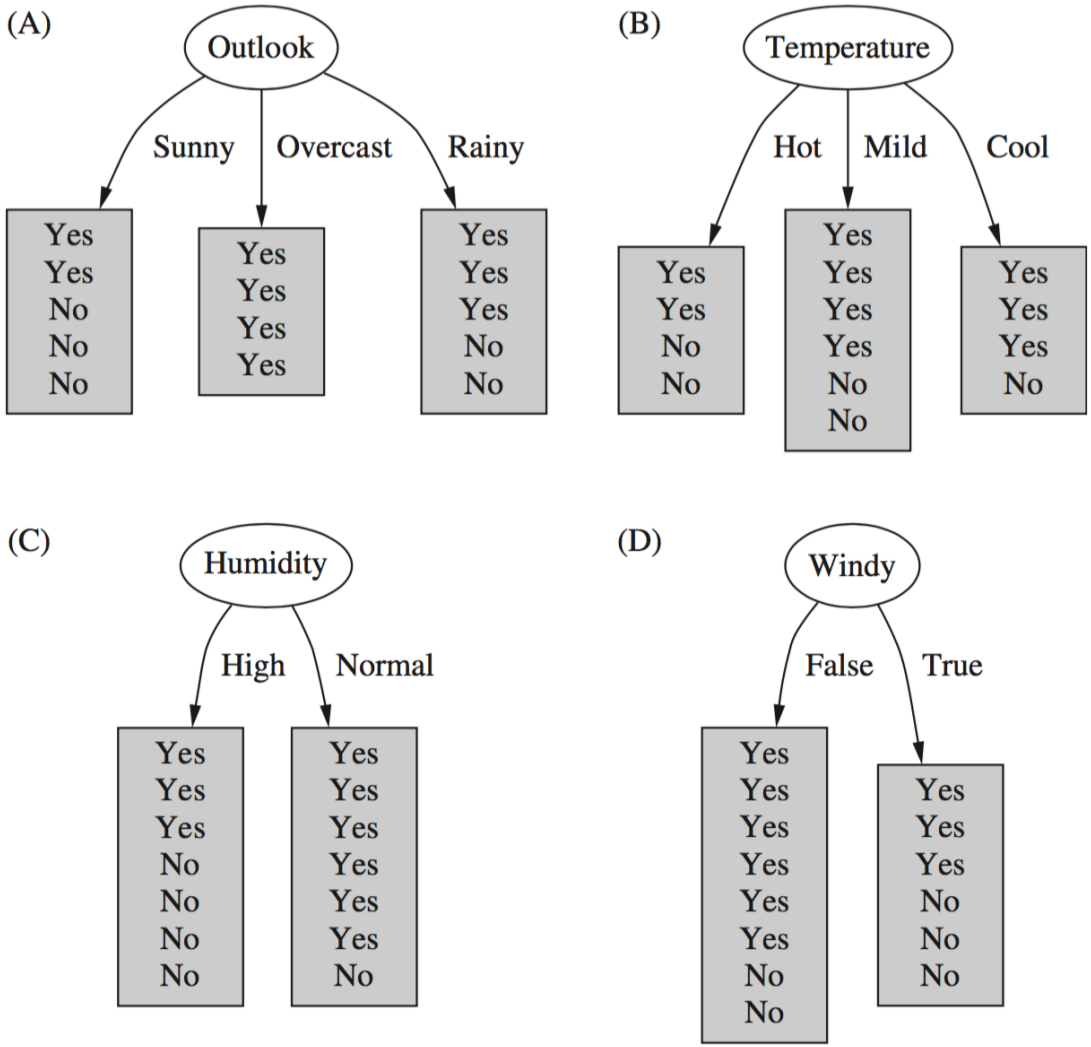

PL<2.5 ?
PL≥2.5 ?
Setosa
PL≥4.9 ?
PL≥4.9 ?
Virginica
PW≥1.6 ?
PW<1.6 ?
Virginica
Versicolor


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

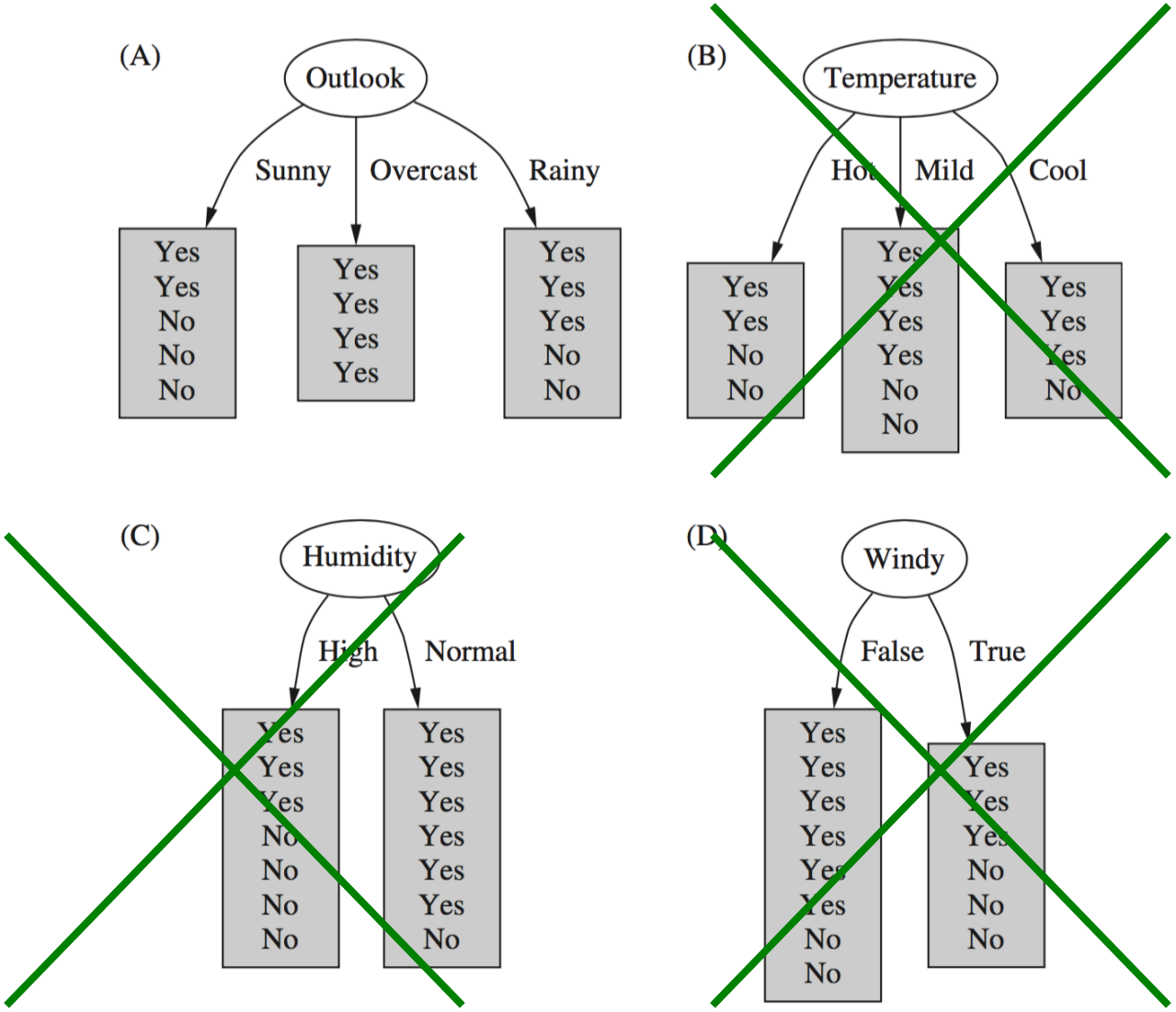
Czy dziś grać w golfa?

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

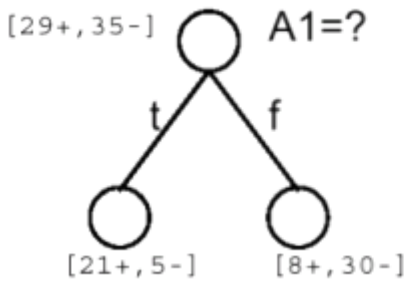
wybór predyktora (1)

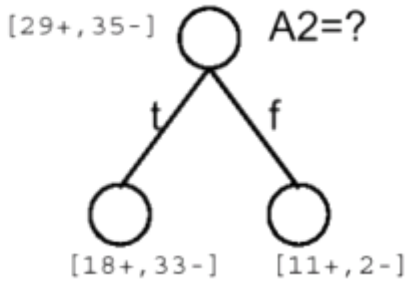


wybór predyktora (2)



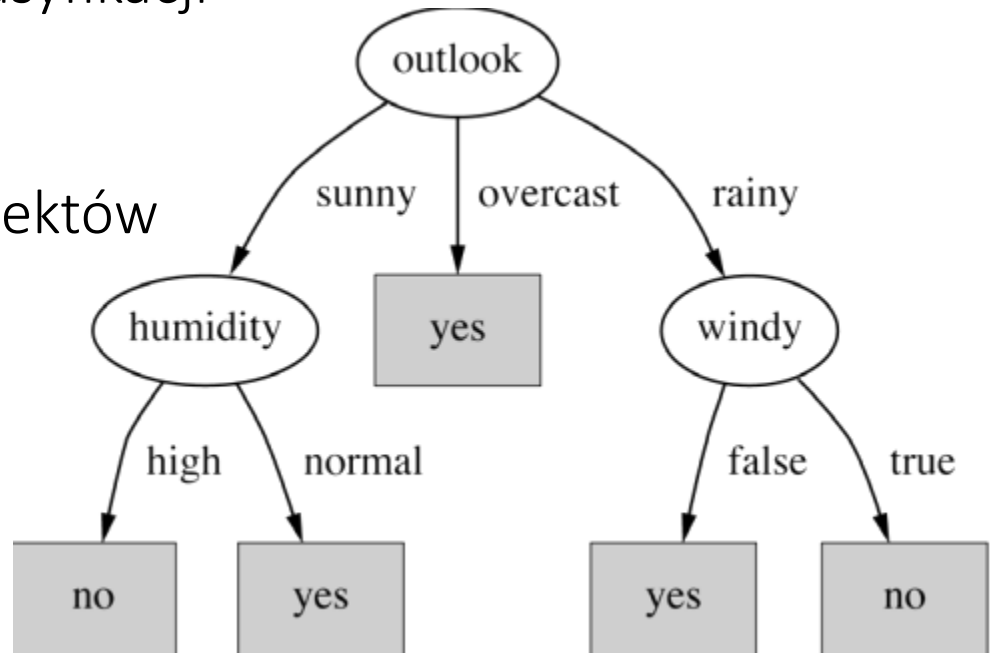
wybór predyktora (3)



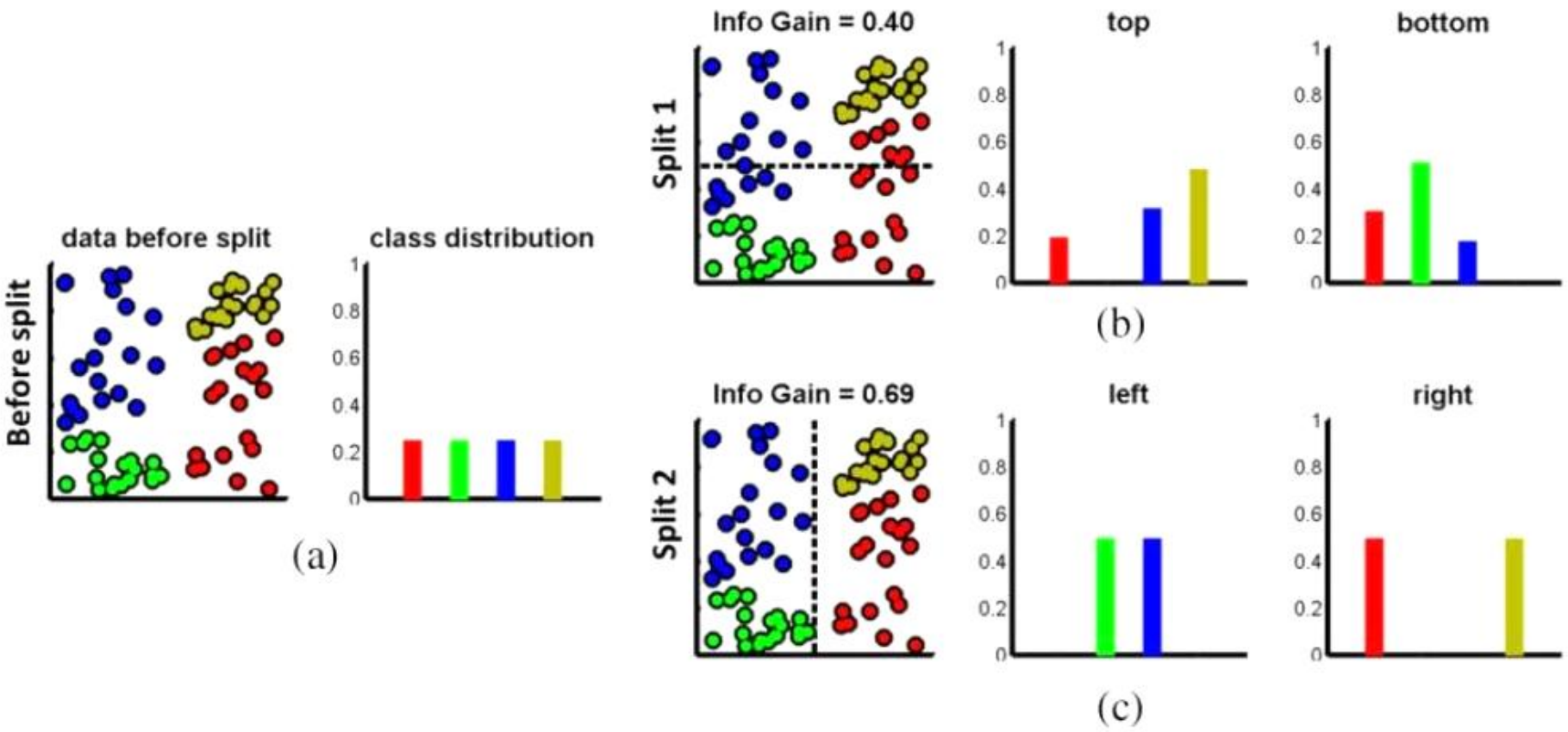


miara (nie)czystości:
- information gain,
- entropy
- Gini Index
- χ^2

- 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

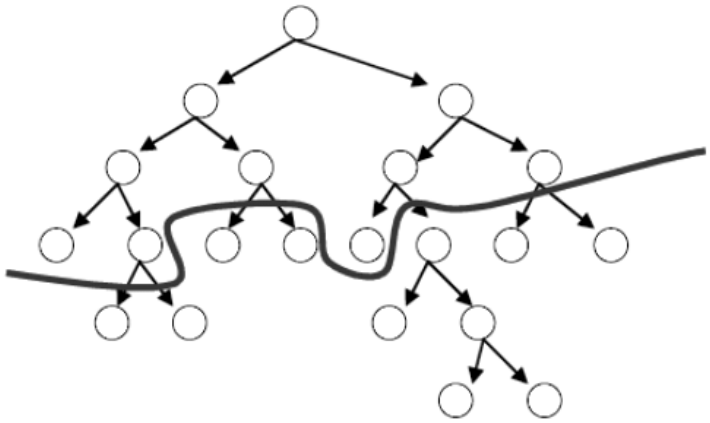
Obcinanie drzewa (1)

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



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.

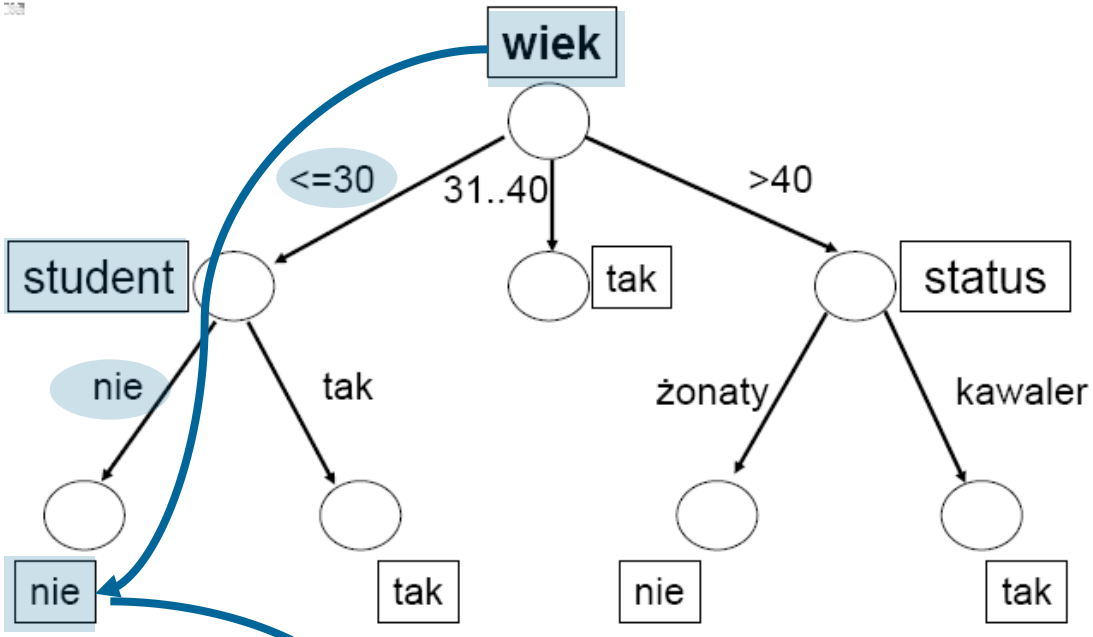
– przycinanie drzewa po zakończeniu konstrukcji drzewa (*postpruning*)

» 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)

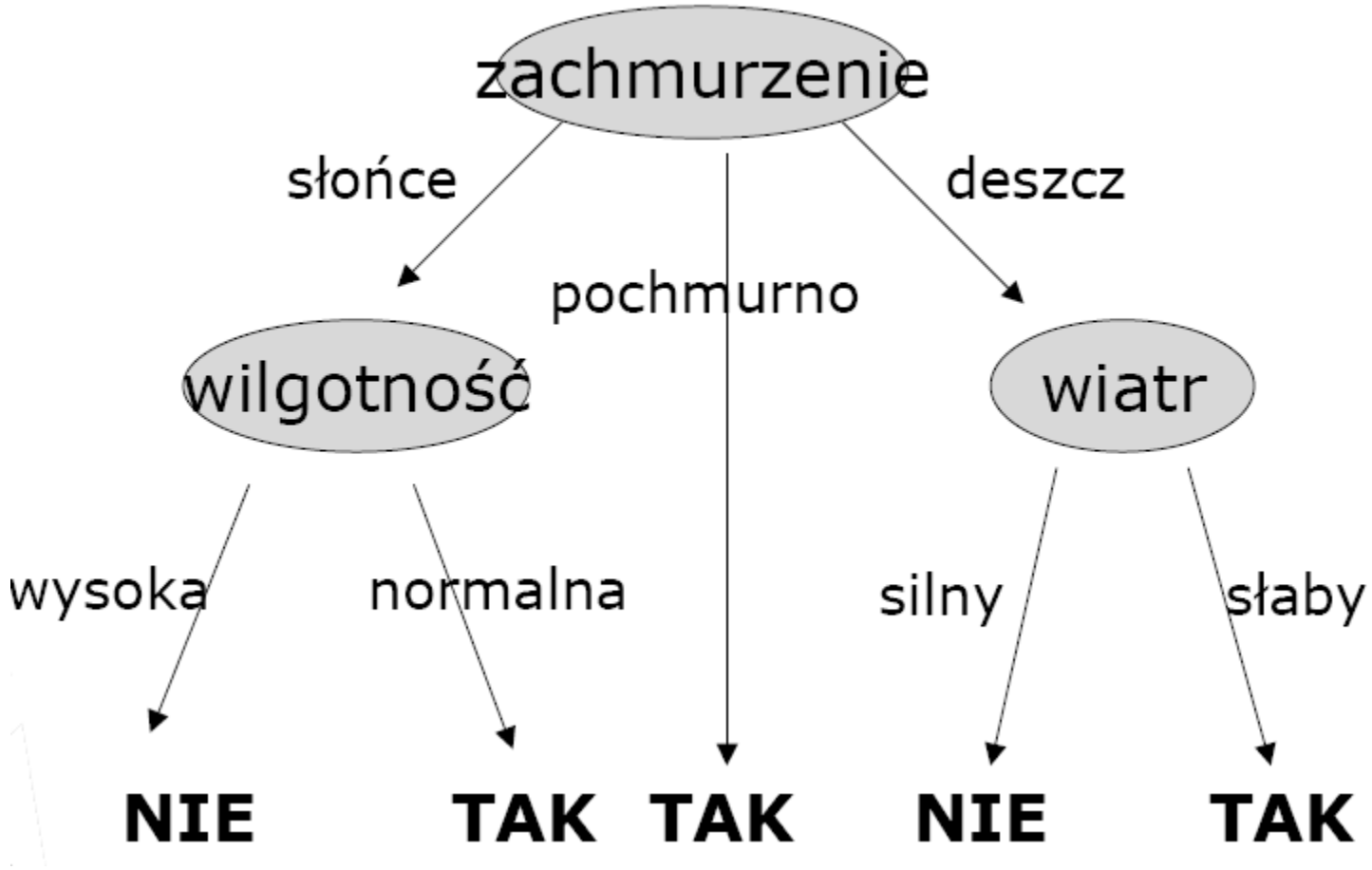


Drzewo decyzyjne można przedstawić w postaci następującego zbioru reguł klasyfikacyjnych:

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



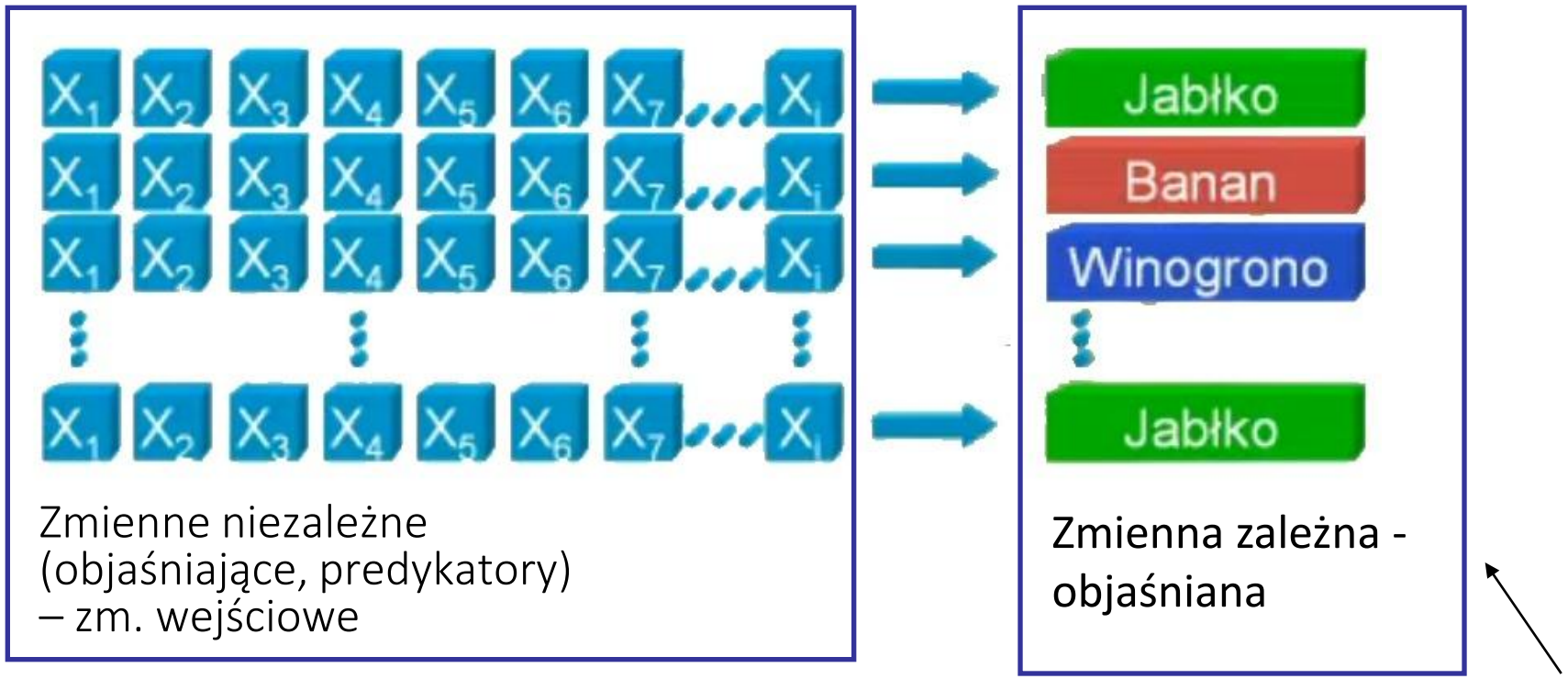
Drzewa – pojęcia podstawowe

- 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)

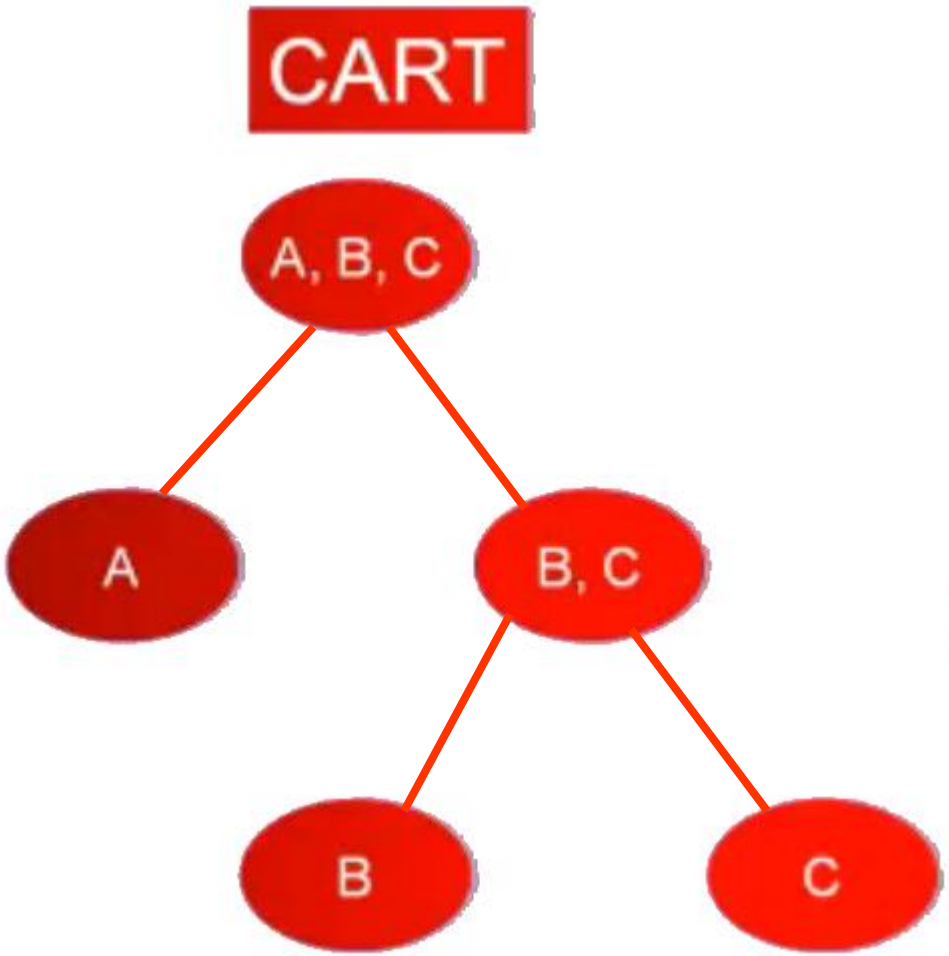
Zmienne w analizie



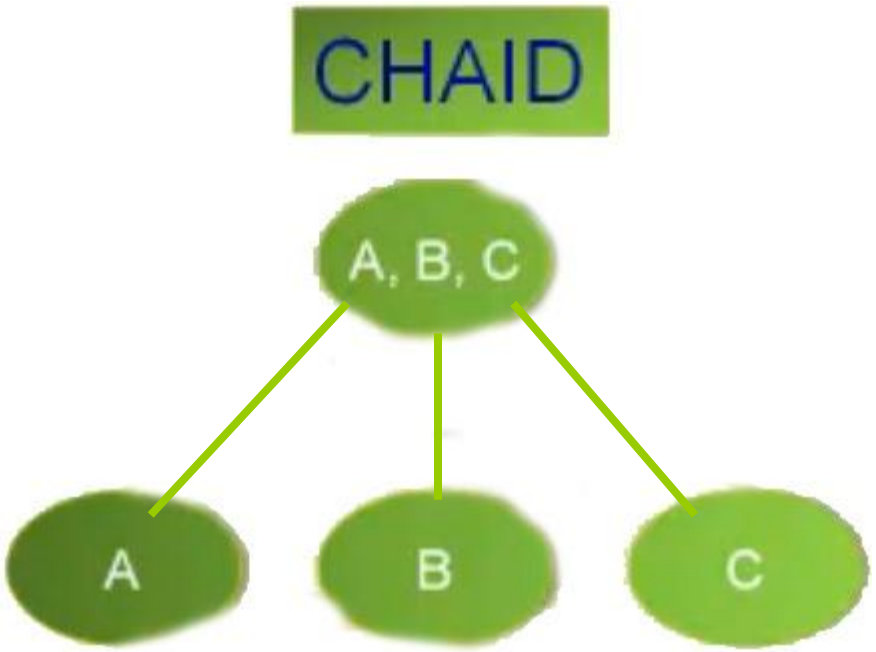
CART (C&RT) i CHAID

Classification & Regression Trees

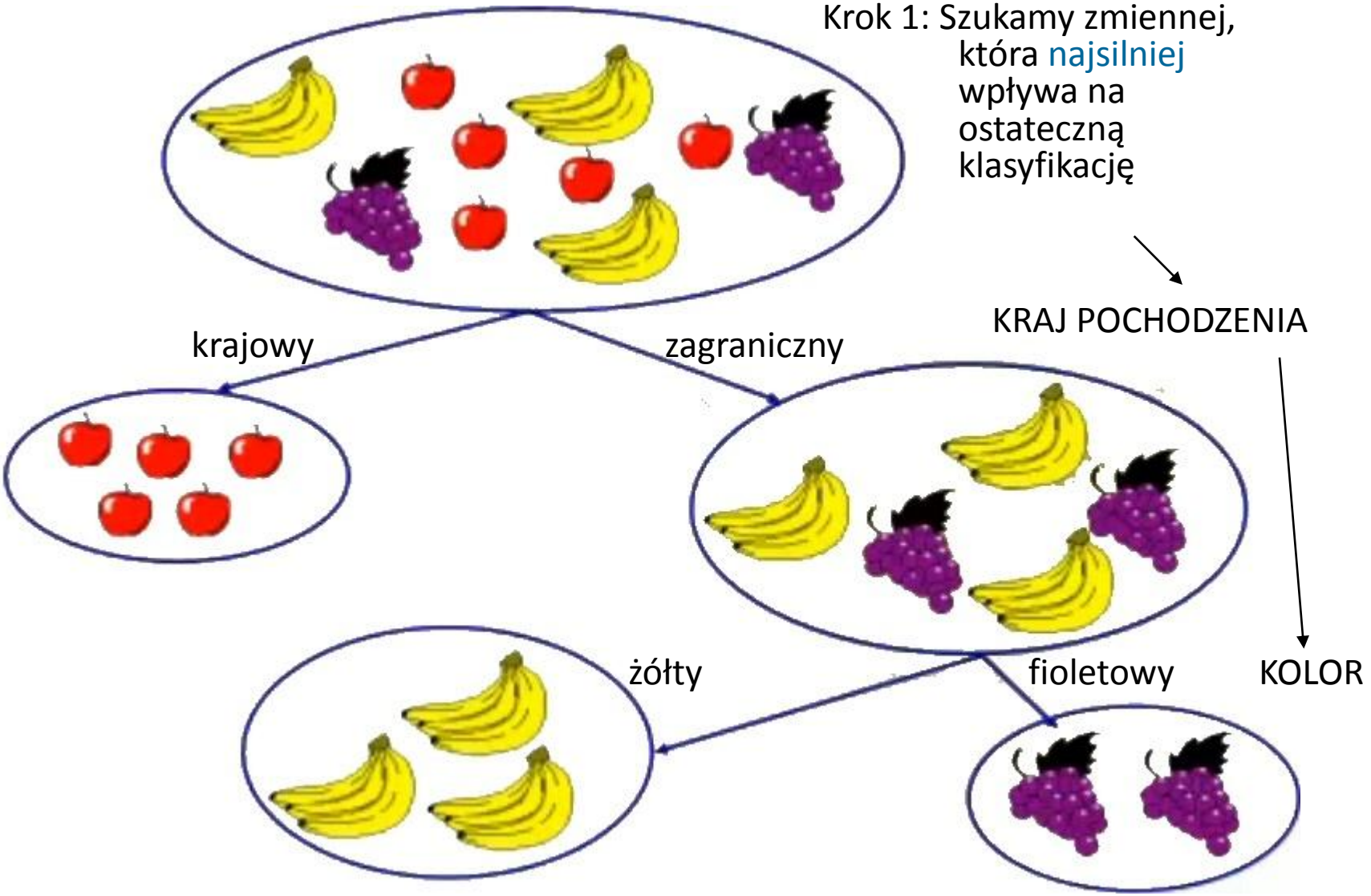
C&RT



Chi-square Automatic Interaction Detection



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

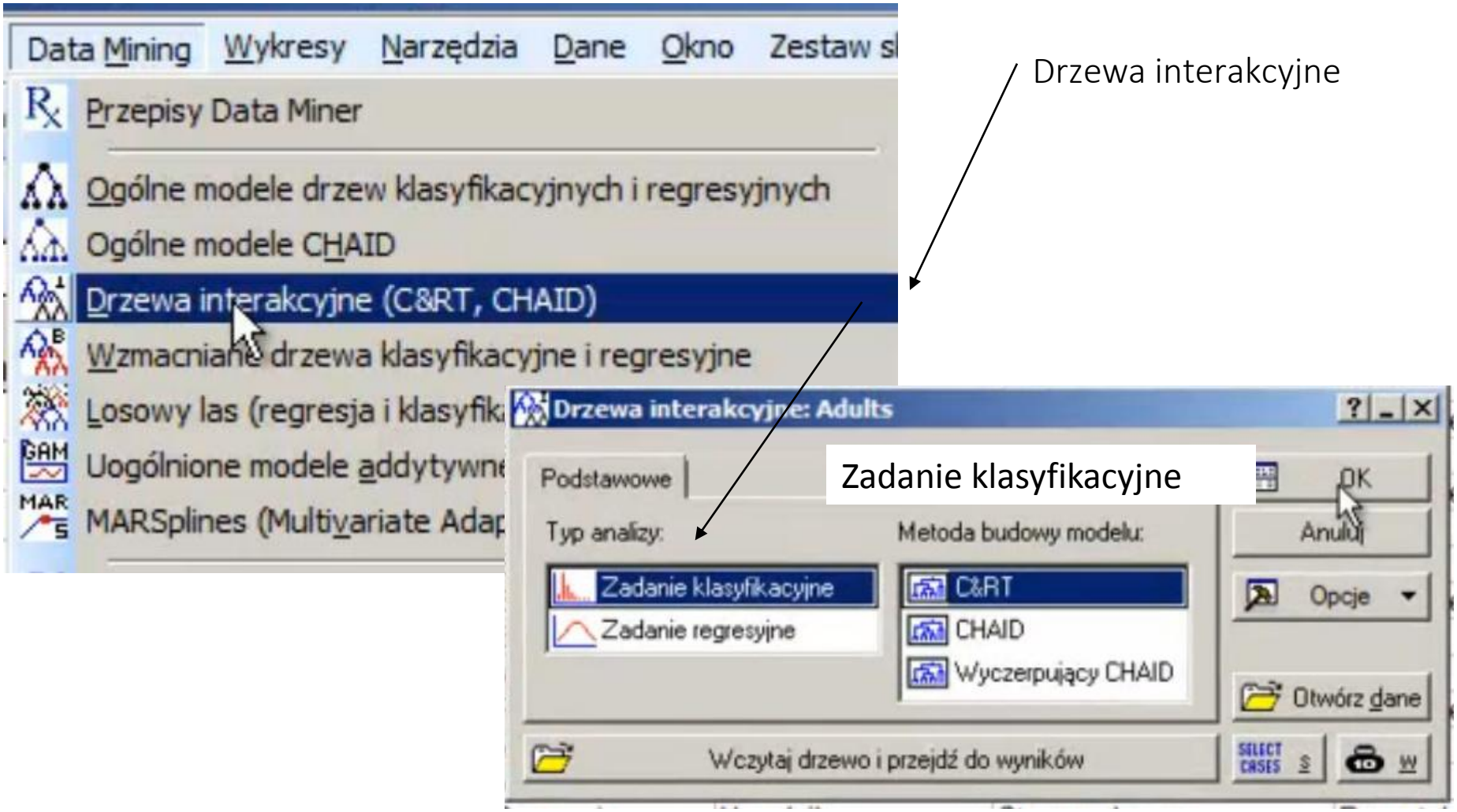
adults (11 zmn. * 32561 prz.)

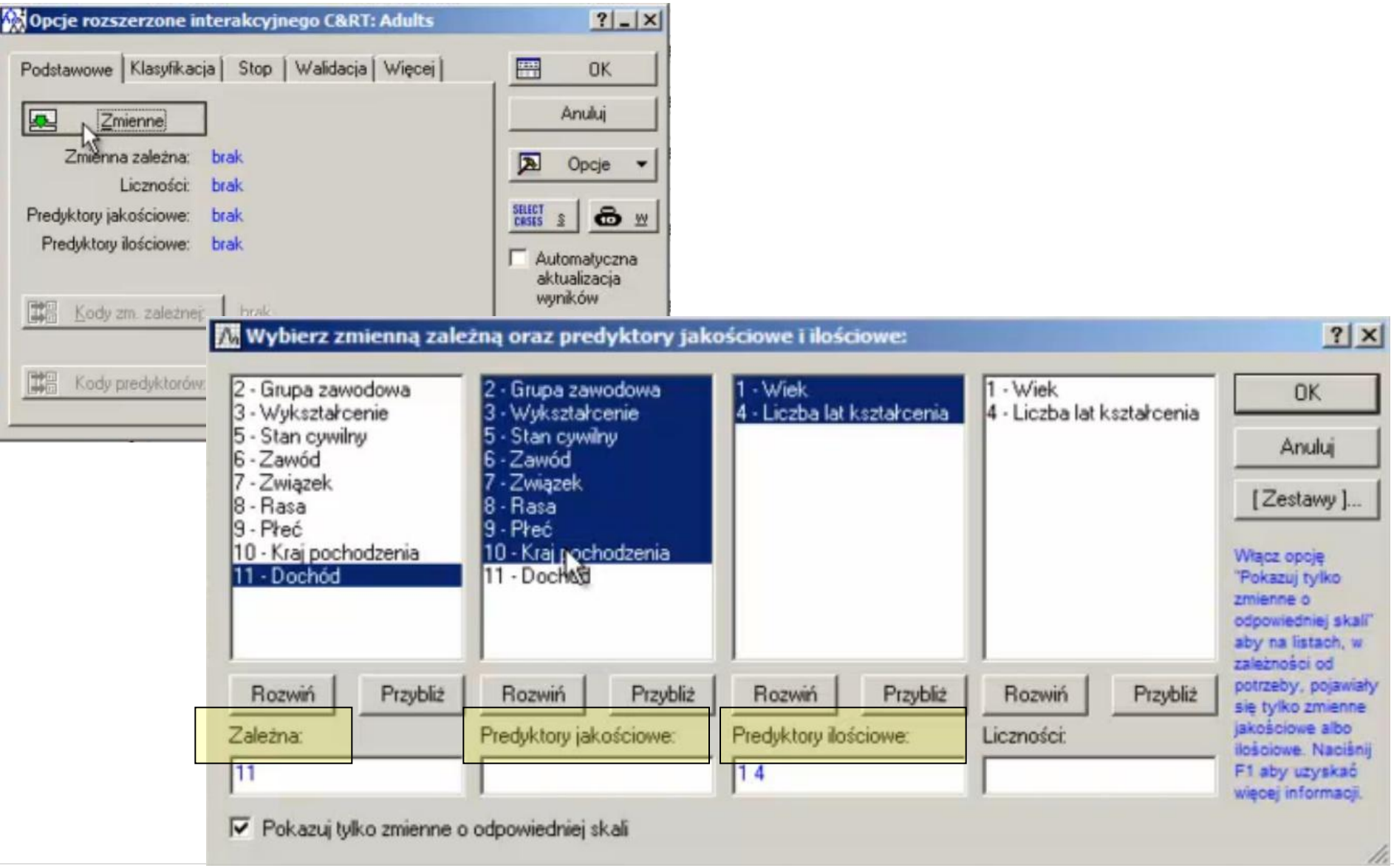
	1 Wiek	2 Grupa zawodowa	3 Wykształcenie	4 Liczba lat kształcenia	5 Stan cywilny	6 Zawód	7 Związek	8 Rasa	9 Płeć
1	25	Firma prywatna	College	10	Rozwiedziona/ny	Kierownik/Dyrektor	Stanu wolnego	Biała	Kobieta
2	39	Firma prywatna	Licencjat	12	Separacja	Mechanik	Poza rodziną	Biała	Mężczyzna
3	45	Firma prywatna	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 federalan	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

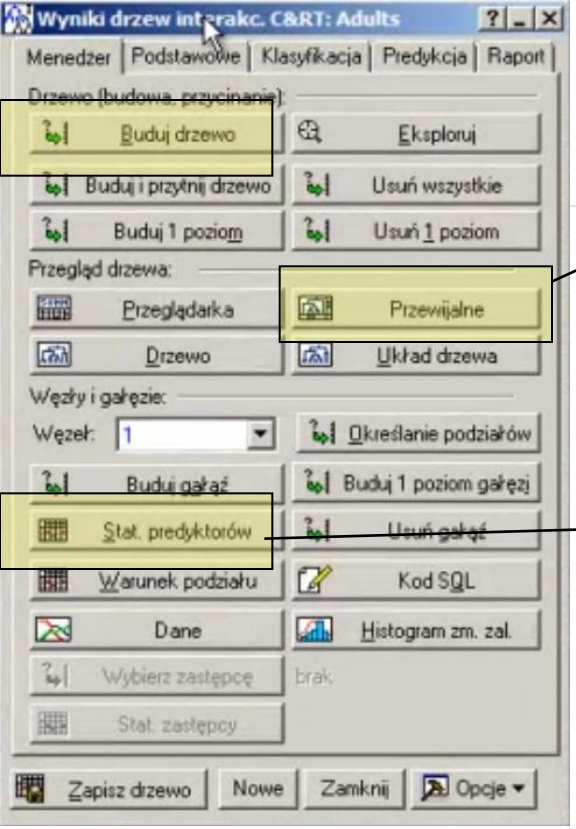
Dochód skategoryzowany do 2 klas:

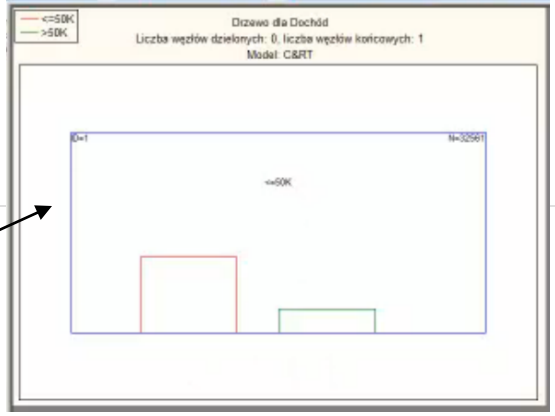
<=50K – poniżej 50 000

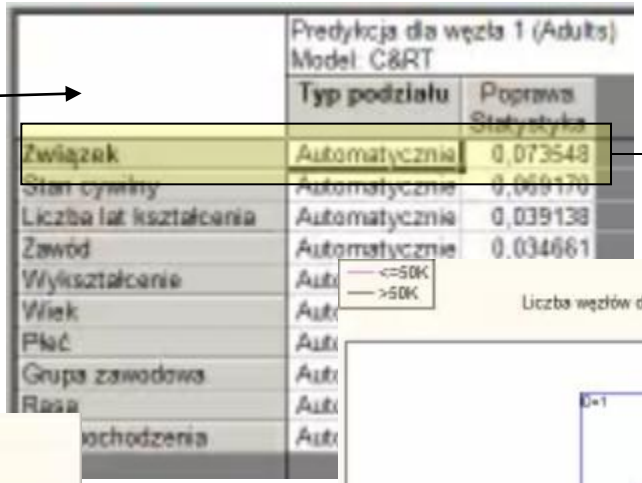
>50K – powyżej 50 000

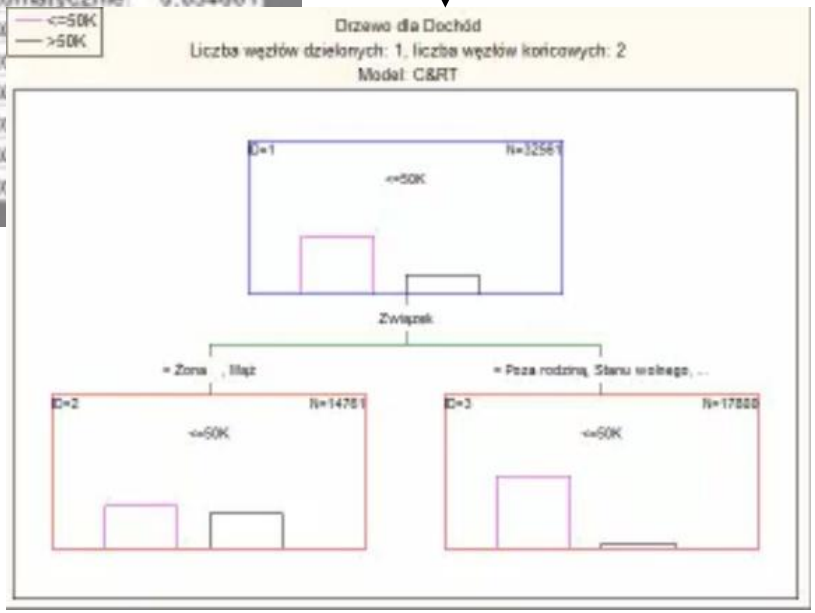

Data Mining
Wykresy
Narzędzia
Dane
Okno
Przepisy Data Miner
Ogólne modele drzew klasyfikacyjnych i regresyjnych
Ogólne modele CHAID
Drzewa interakcyjne (C&RT, CHAID)
Wzmacniane drzewa klasyfikacyjne i regresyjne
Losowy las (regresja i klasyfik
Uogólnione modele addytywne
MARSplines (Multivariate Adap
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
SELECT CASES

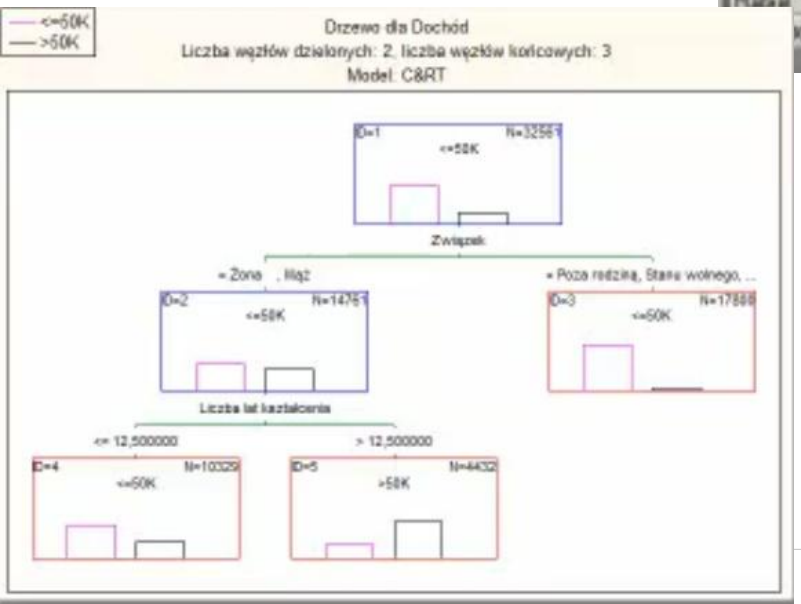

Opcje rozszerzone interakcyjnego C&RT: Adults
Podstawowe
Klasyfikacja
Stop
Walidacja
Więcej
OK
Anuluj
Zmienne
Zmienna zależna: brak
Liczności: brak
Predyktory jakościowe: brak
Predyktory ilościowe: brak
Opcje
Automatyczna aktualizacja wyników
Kody zm. zależnej:
Kody predyktorów:
Wybierz zmienną zależną 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
OK
Anuluj
[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

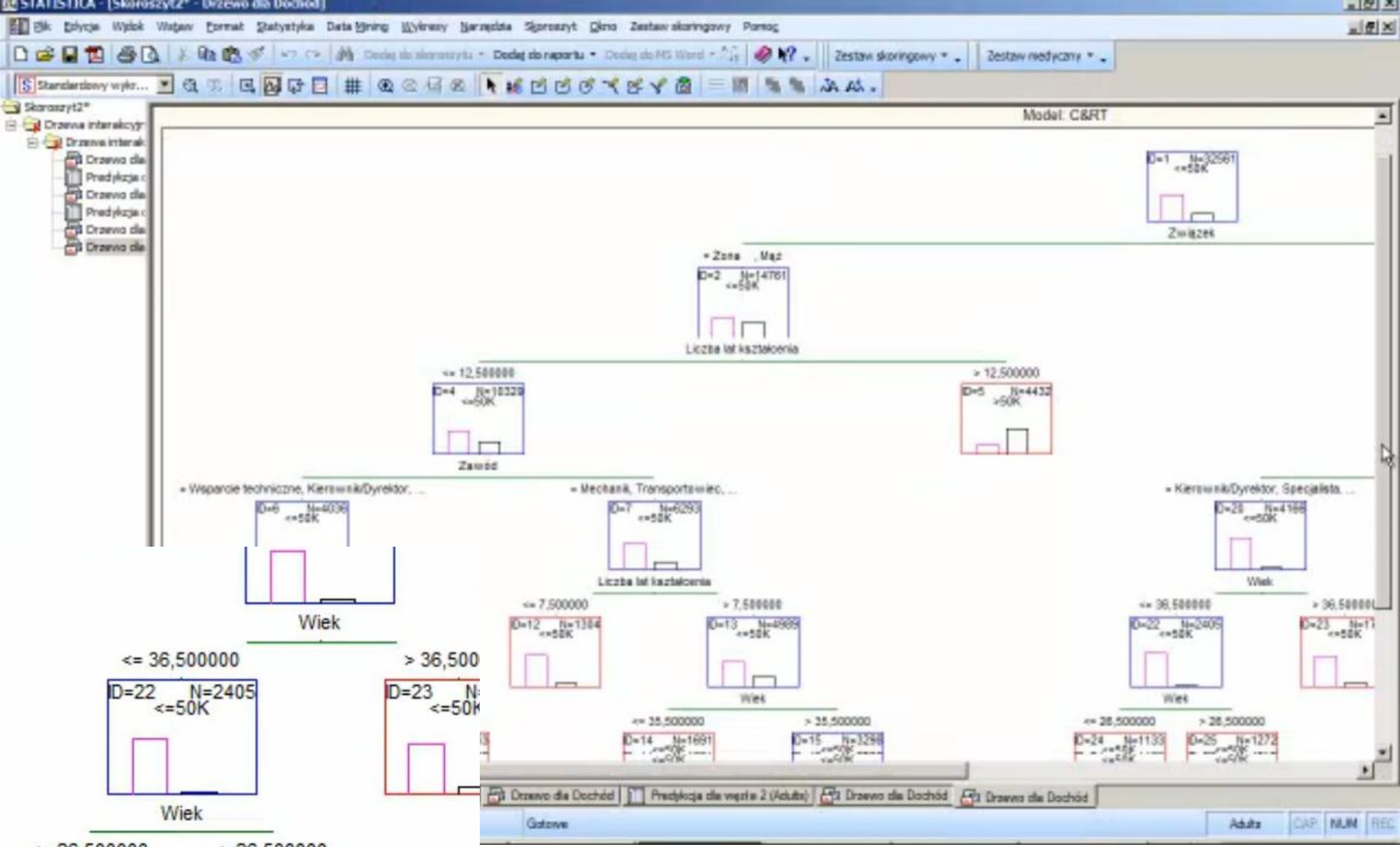

Wyniki drzew interakc. C&RT: Adults
Menedżer
Podstawowe
Klasyfikacja
Predykcja
Raport
Drzewo (budowa, przycinanie):
Buduj drzewo
Eksploruj
Buduj i przytnij drzewo
Usuń wszystkie
Buduj 1 poziom
Usuń 1 poziom
Przegląd drzewa:
Przeglądarka
Przewijalne
Drzewo
Układ drzewa
Węzły i gałęzie:
Węzeł: 1
Określanie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

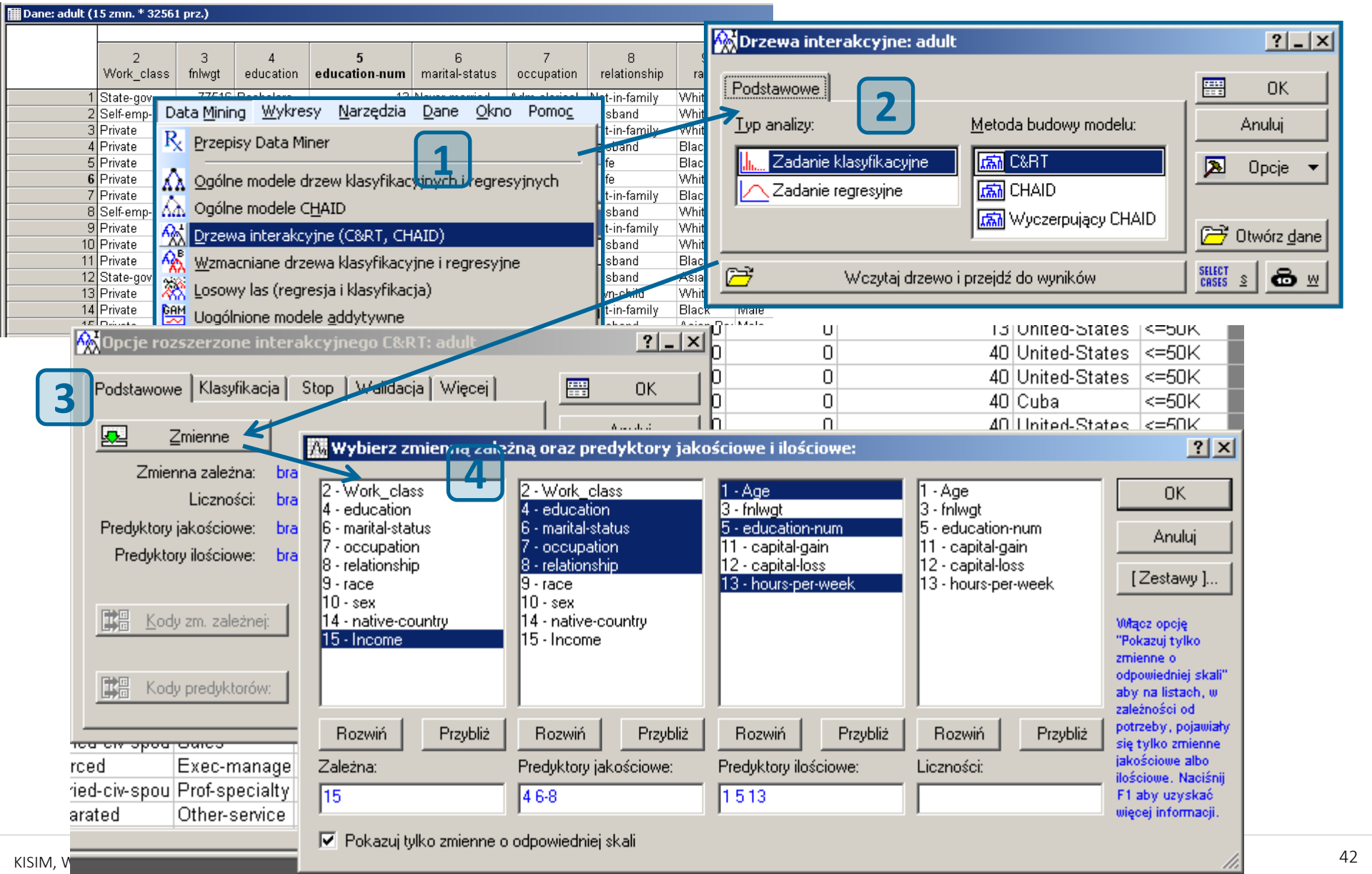

Predykcja dla węzła 1 (Adults)
Model: C&RT
Typ podziału
Poprawa
Statystyka
Związek
Automatycznie
0,073548
Stan cywilny
Automatycznie
0,069170
Liczba lat kształcenia
Automatycznie
0,039138
Zawód
Automatycznie
0,034661
Wykształcenie
Wiek
Płeć
Grupa zawodowa
Rasa
Pochodzenia

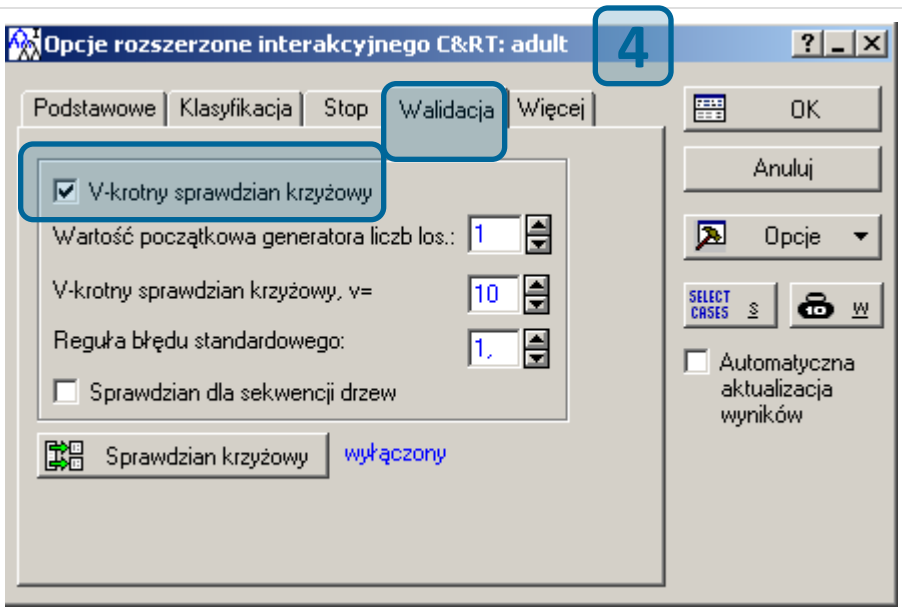

<=50K
>50K
Drzewo dla Dochód
Liczba węzłów dzielonych: 1, liczba węzłów końcowych: 2
Model: C&RT
ID=1
N=32561
<=50K
Związek
= Żona , Mąż
= Poza rodziną, Stanu wolnego, ...
ID=2
N=14761
<=50K
ID=3
N=17800
<=50K

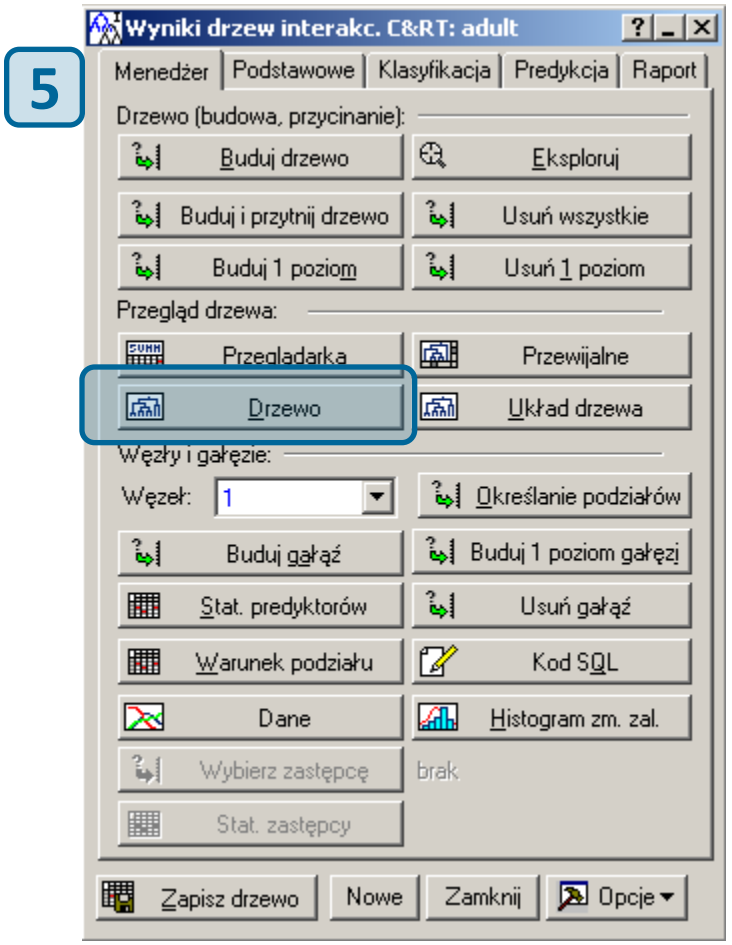

<=50K
>50K
Drzewo dla Dochód
Liczba węzłów dzielonych: 2, liczba węzłów końcowych: 3
Model: C&RT
ID=1
N=32561
<=50K
Związek
= Żona , Mąż
= Poza rodziną, Stanu wolnego, ...
ID=2
N=14761
<=50K
ID=3
N=17800
<=50K
Liczba lat kształcenia
<= 12,500000
> 12,500000
ID=4
N=10329
<=50K
ID=5
N=4432
>50K

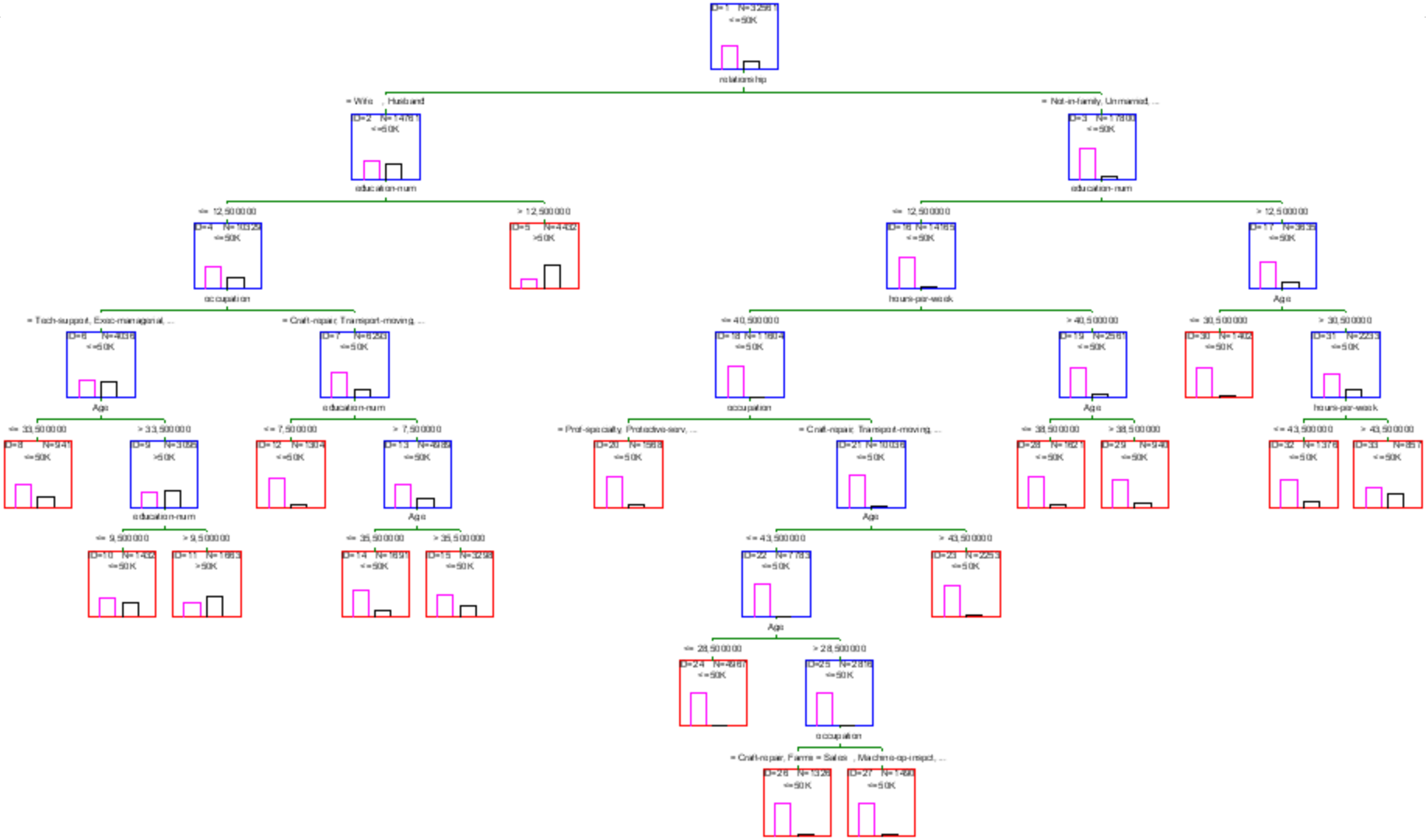

Wiek
<= 36,500000
> 36,500
ID=22 N=2405
<=50K
Wiek
<= 26,500000
> 26,500000
ID=24 N=1133
<=50K
ID=25 N=1272
<=50K

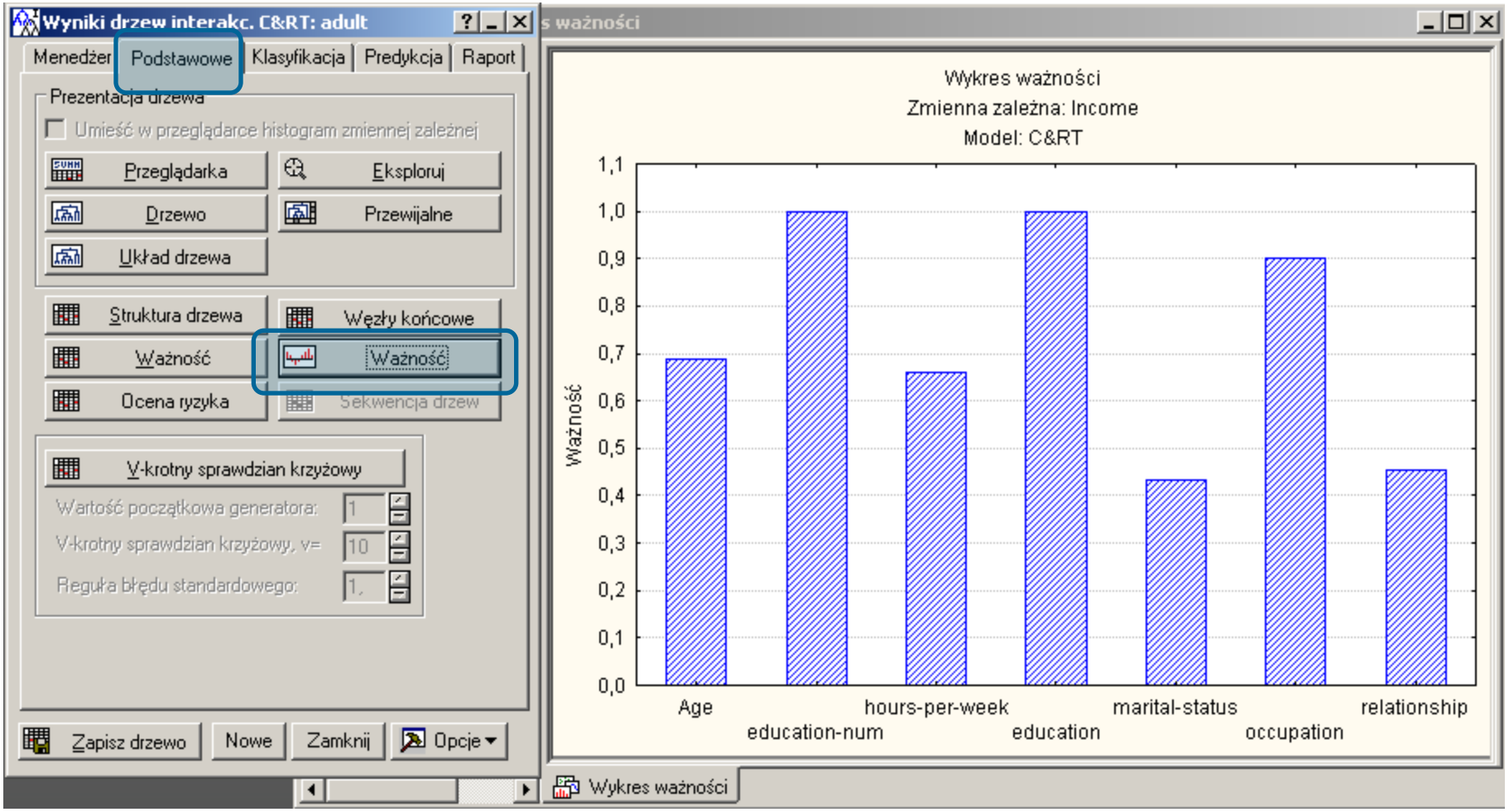
STATISTICA - przykład drzewa C&RT (inny dobór zmiennych)

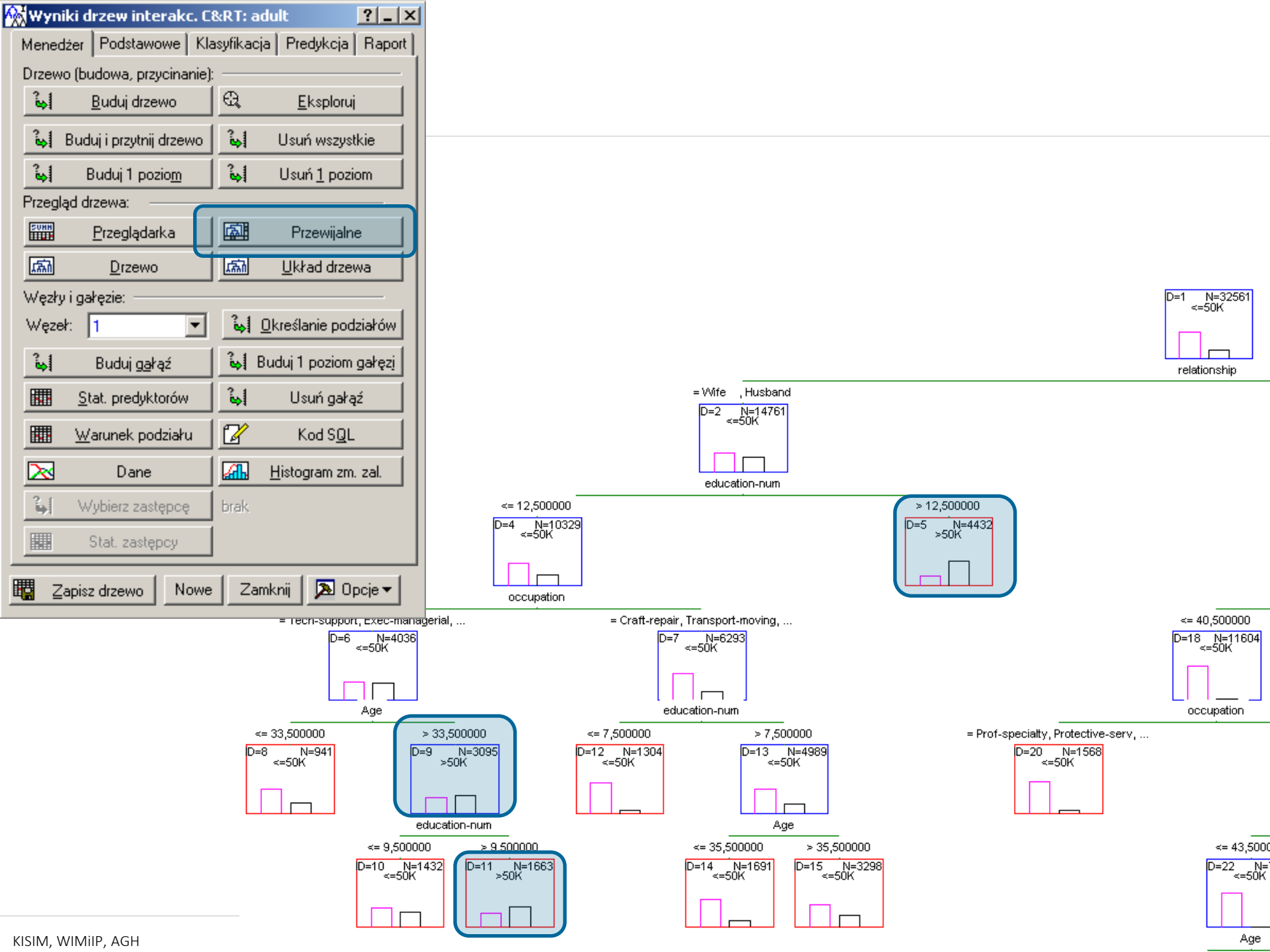


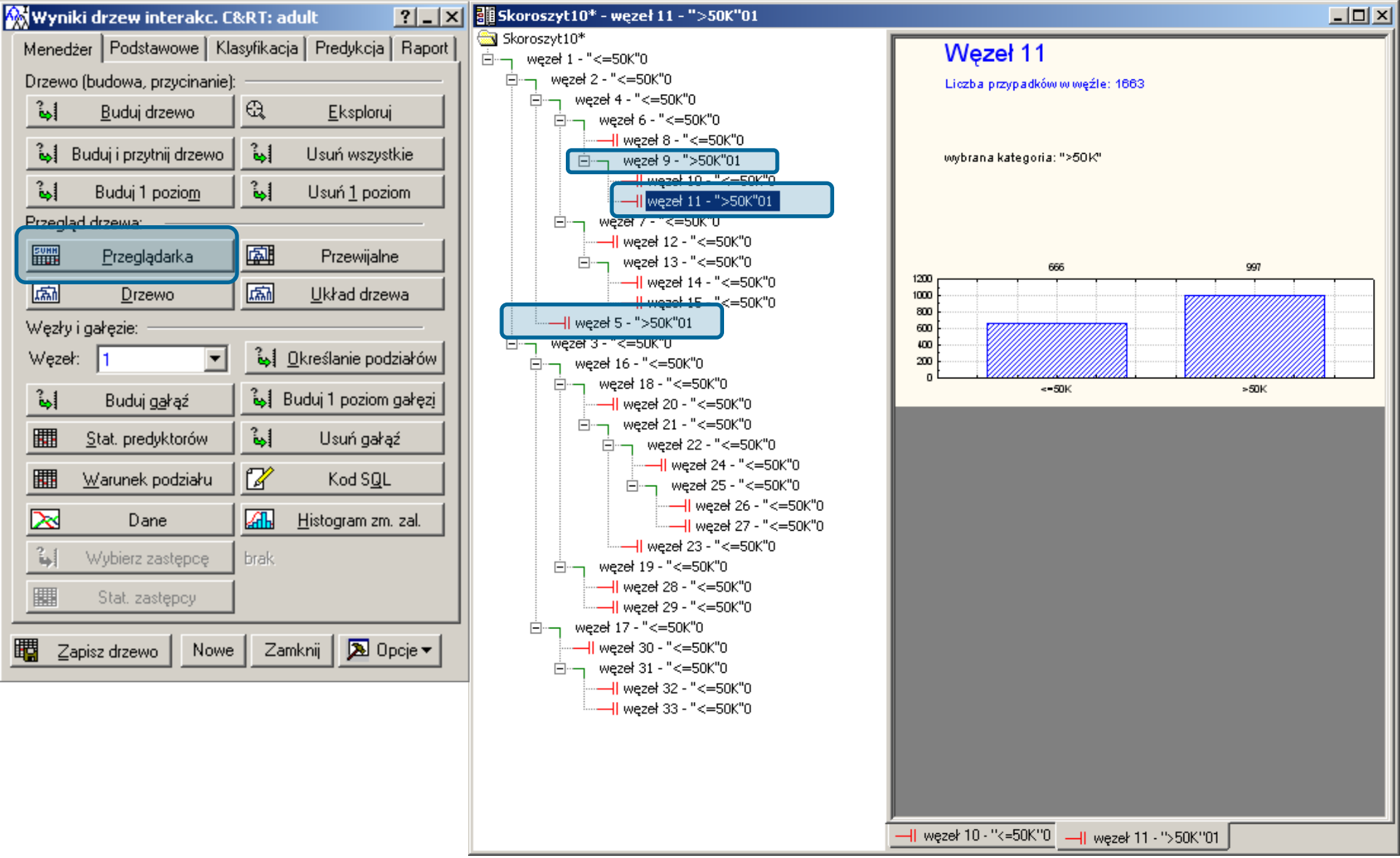

4
Opcje rozszerzone interakcyjnego C&RT: adult
Podstawowe
Klasyfikacja
Stop
Walidacja
Więcej
V-krotny sprawdzian krzyżowy
Wartość początkowa generatora liczb los.:
1
V-krotny sprawdzian krzyżowy, v=
10
Reguła błędu standardowego:
1,
Sprawdzian dla sekwencji drzew
Sprawdzian krzyżowy
wyłączony
OK
Anuluj
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
<= 12.500000
> 12.500000
<= 12.500000
> 12.500000
occupation
hours-per-week
Age
= Tech-support, Exec-managerial, ...
= Craft-repair, Transport-moving, ...
<= 40.500000
> 40.500000
<= 30.500000
> 30.500000
Age
education-num
occupation
Age
Age
hours-per-week
<= 33.500000
> 33.500000
<= 7.500000
> 7.500000
= Prof-specialty, Protective-serv, ...
= Craft-repair, Transport-moving, ...
<= 38.500000
> 38.500000
<= 43.500000
> 43.500000
education-num
Age
Age
<= 9.500000
> 9.500000
<= 35.500000
> 35.500000
<= 43.500000
> 43.500000
<= 28.500000
> 28.500000
occupation
= Craft-repair, Farms = Sales , Machine-op-inspct, ...

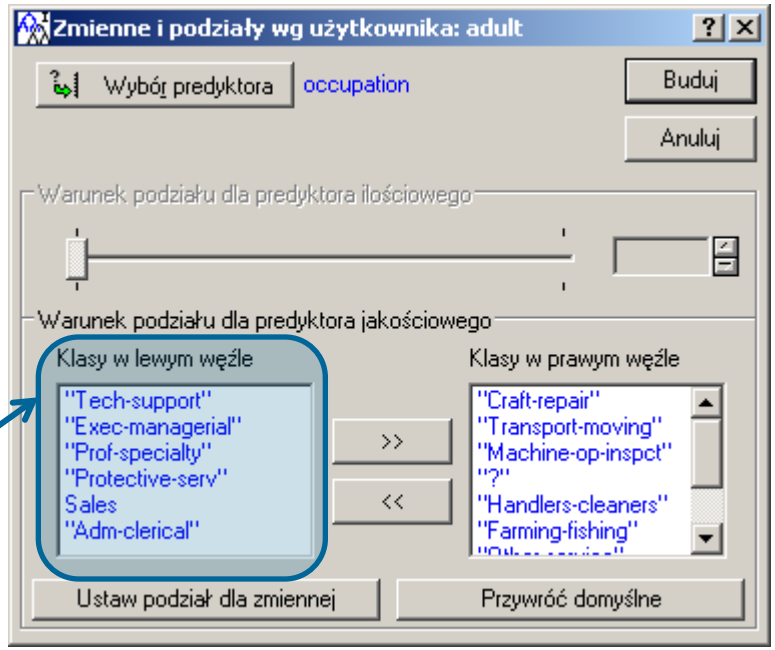

Wyniki drzew interakc. C&RT: adult
Menedżer
Podstawowe
Klasyfikacja
Predykcja
Raport
Prezentacja drzewa
Umieść w przeglądarce histogram zmiennej zależnej
Przeglądarka
Eksploruj
Drzewo
Przewijalne
Układ drzewa
Struktura drzewa
Węzły końcowe
Ważność
Ważność
Ocena ryzyka
Sekwencja drzew
V-krotny sprawdzian 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

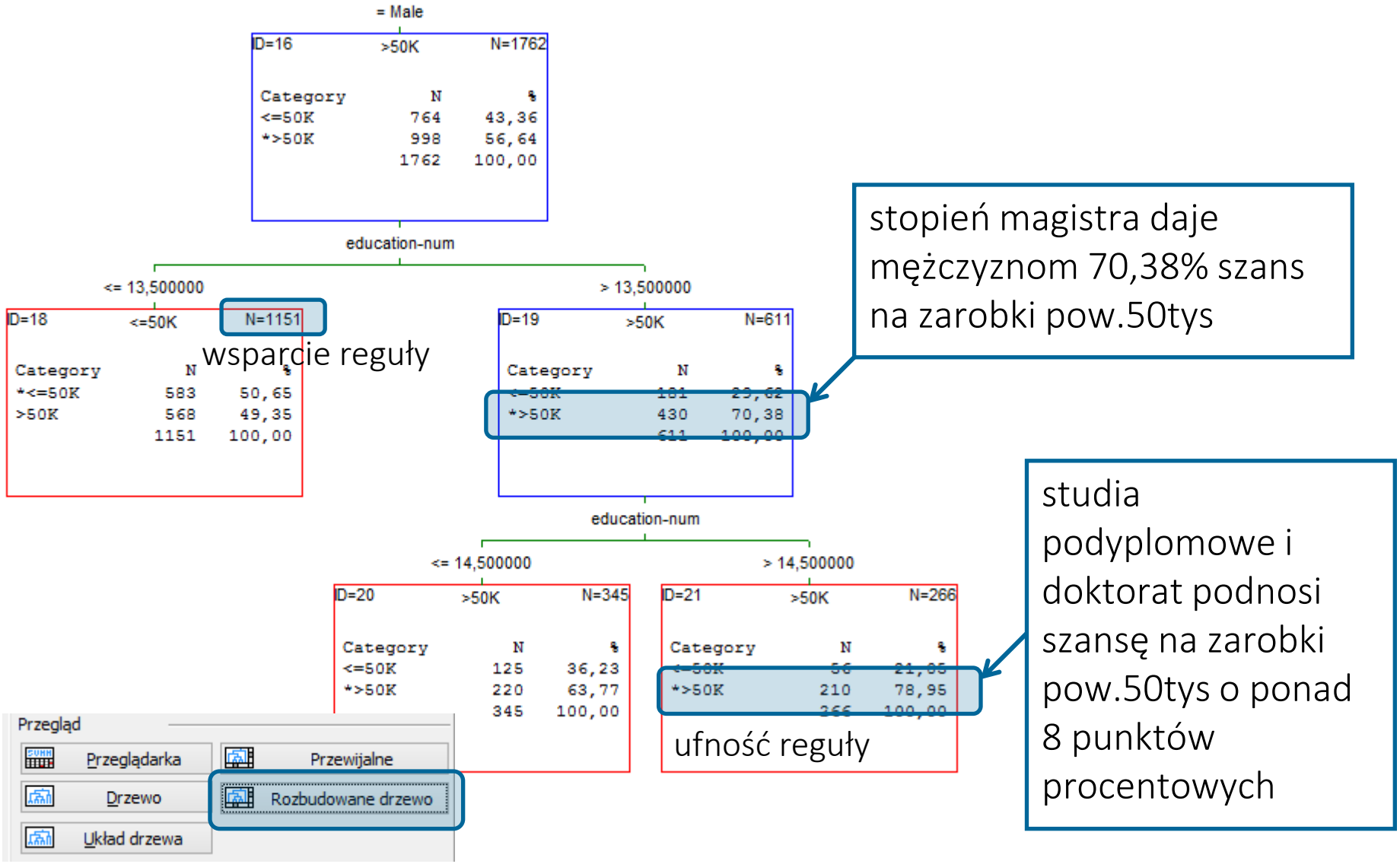

Wyniki drzew interakc. C&RT: adult
Menedżer
Podstawowe
Klasyfikacja
Predykcja
Raport
Drzewo (budowa, przycinanie):
Buduj drzewo
Eksploruj
Buduj i przytnij drzewo
Usuń wszystkie
Buduj 1 poziom
Usuń 1 poziom
Przegląd drzewa:
Przeglądarka
Przewijalne
Drzewo
Układ drzewa
Węzły i gałęzie:
Węzeł:
1
Określanie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

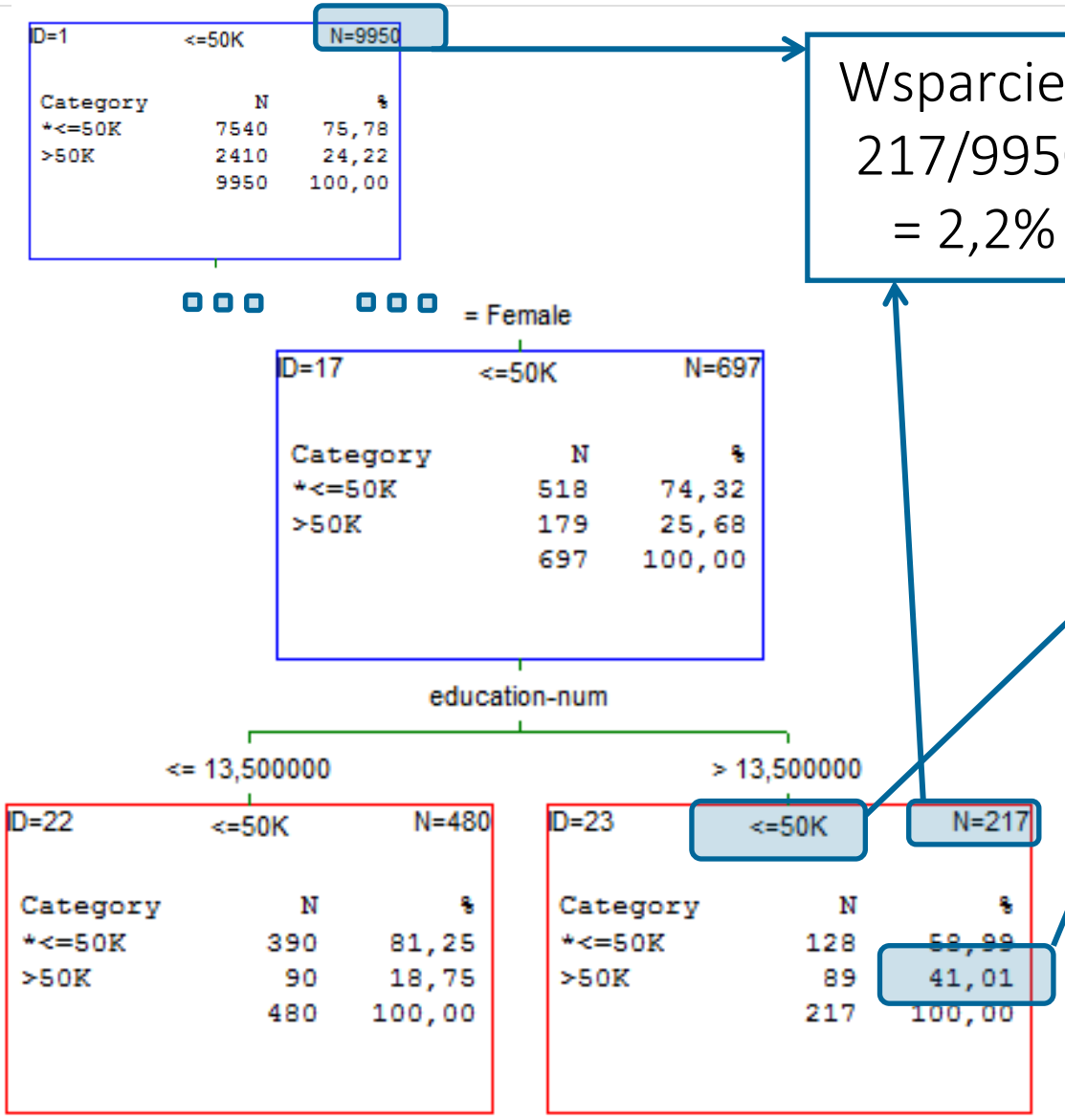
- Jeżeli osoba pozostaje w związku małżeńskim i jej liczba lat edukacji przekracza 12,5 roku, wtedy jej dochód prawdopodobnie przekracza 50 000 $ (węzeł ID5)
(z prawdopodobieństwem… 72%)
- Jeżeli osoba pozostaje w związku małżeńskim, jej liczba lat edukacji nie przekracza 12,5 roku, wykonuje zawód… oraz 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%)



Pewność reguł (ufność reguły, dokładność)



Wsparcie i Ufność



jest bardzo mało kobiet wysoko wykształconych

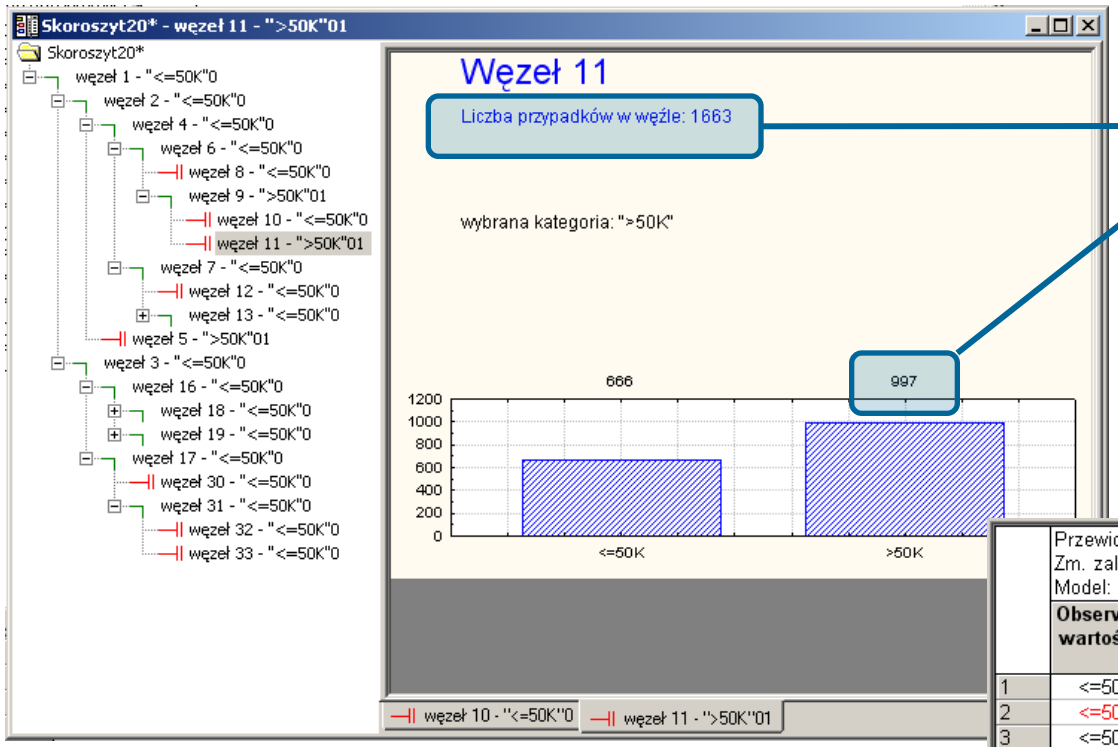
Wsparcie = 217/9950 = 2,2%

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

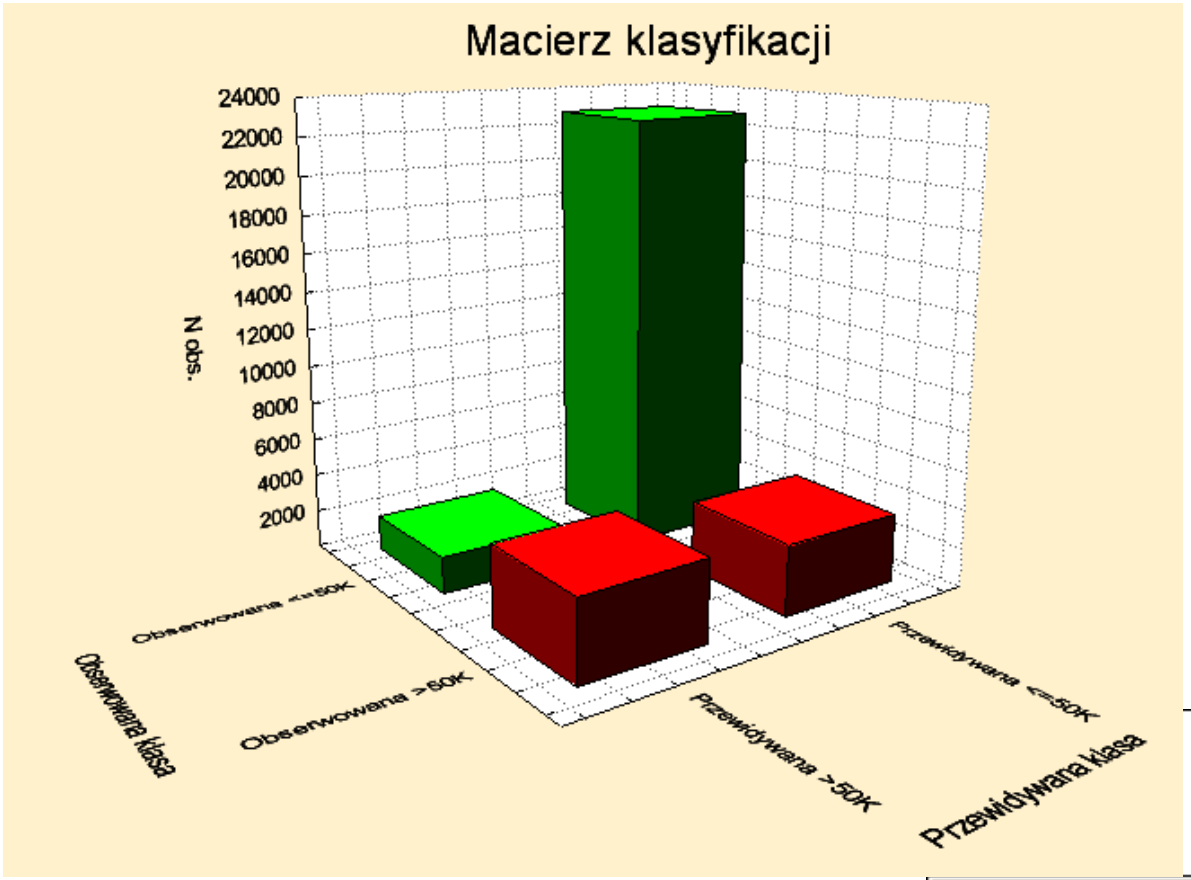
Macierz klasyfikacji (adult) | Macierz klasyfikacji (adult) | Przewidywane (adult) | Wykres przyros

wsparcie reguły:

$N_{węzła} / N_{zbioru}$

ID11: 1633/32561=5%

Macierz klasyfikacji



Ile razy model się pomylił? Pojęcie „kosztu”

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

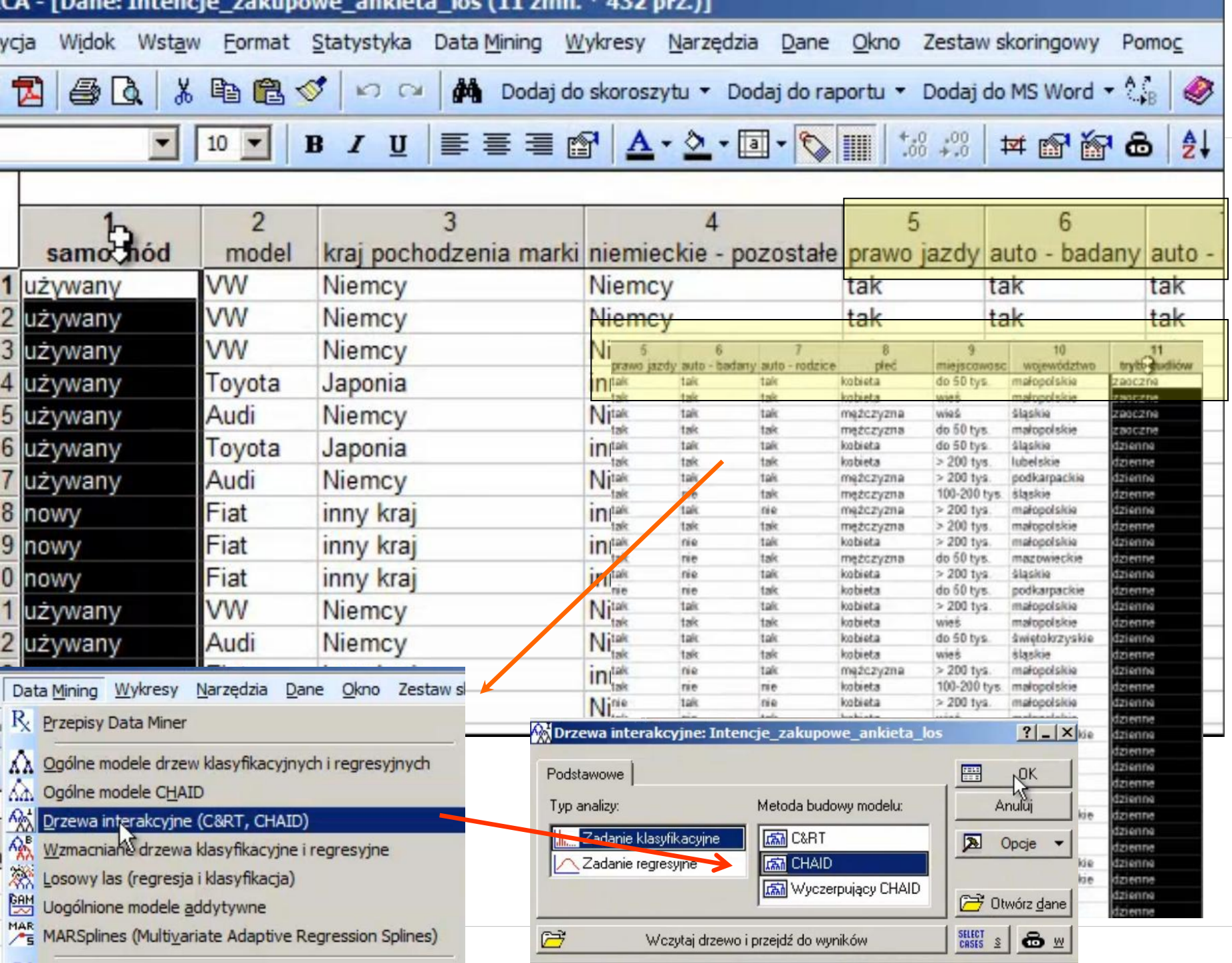
FN

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



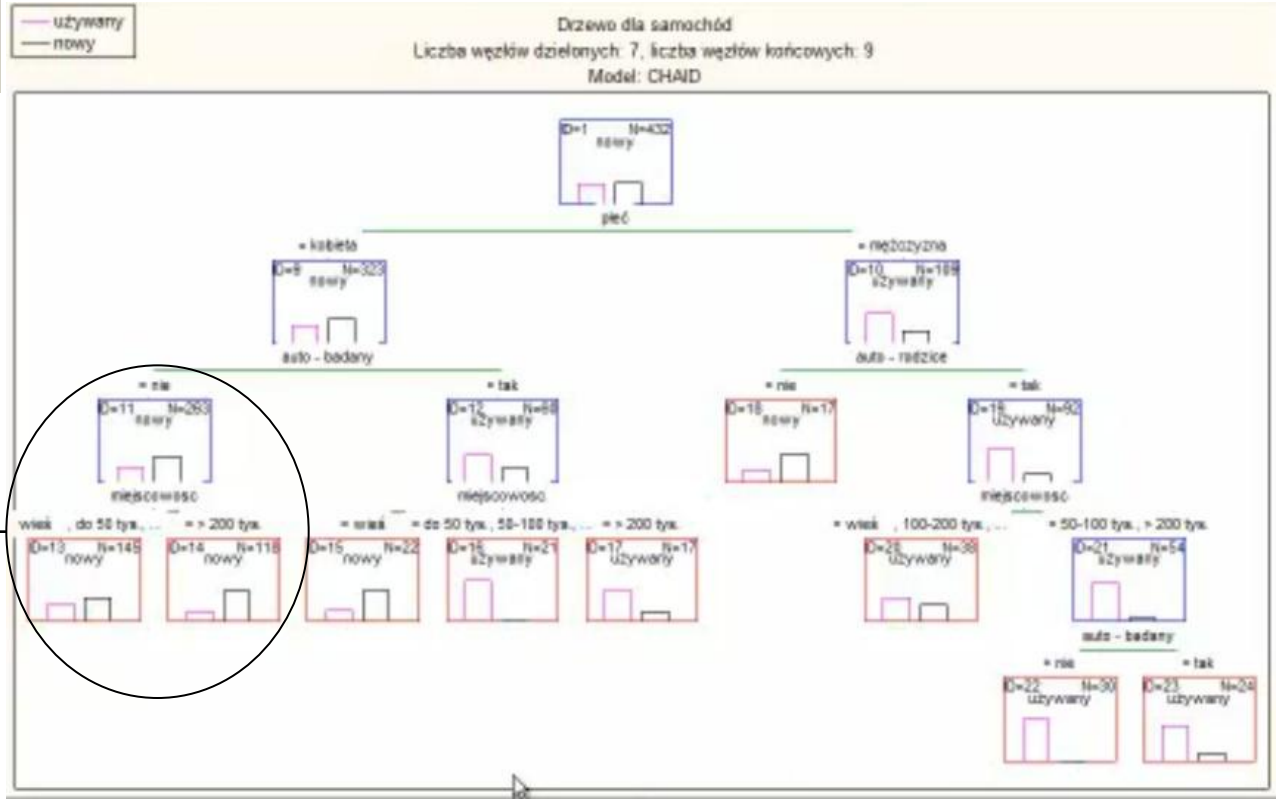
1 samochód	2 model	3 kraj pochodzenia marki	4 niemieckie - pozostałe	5 prawo jazdy	6 auto - badany	auto -
1 używany	VW	Niemcy	Niemcy	tak	tak	tak
2 używany	VW	Niemcy	Niemcy	tak	tak	tak
3 używany	VW	Niemcy	Ni			
4 używany	Toyota	Japonia	in			
5 używany	Audi	Niemcy	Ni			
6 używany	Toyota	Japonia	in			
7 używany	Audi	Niemcy	Ni			
8 nowy	Fiat	inny kraj	in			
9 nowy	Fiat	inny kraj	in			
0 nowy	Fiat	inny kraj	in			
1 używany	VW	Niemcy	Ni			
2 używany	Audi	Niemcy	Ni			

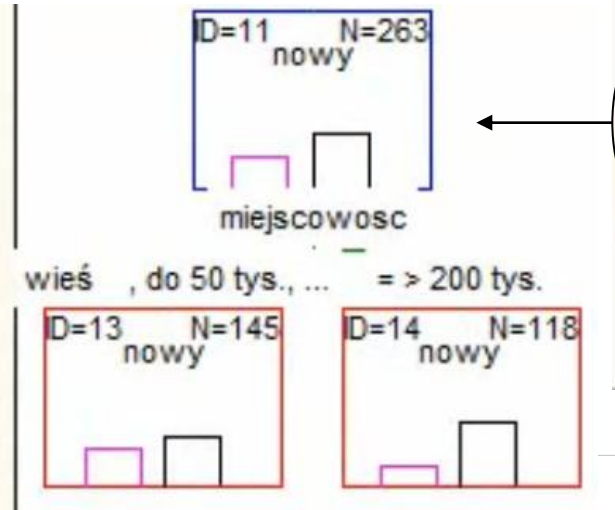
5 prawo jazdy	6 auto - badany	7 auto - rodzice	8 płeć	9 miejscowosc	10 województwo	11 tryb budków
tak	tak	tak	kobieta	do 50 tys.	małopolskie	zaoczne
tak	tak	tak	kobieta	wieś	małopolskie	zaoczne
tak	tak	tak	mężczyzna	wieś	śląskie	zaoczne
tak	tak	tak	mężczyzna	do 50 tys.	małopolskie	zaoczne
tak	tak	tak	kobieta	do 50 tys.	śląskie	dzienne
tak	tak	tak	kobieta	> 200 tys.	lubelskie	dzienne
tak	tak	tak	mężczyzna	> 200 tys.	podkarpackie	dzienne
tak	nie	tak	mężczyzna	100-200 tys.	śląskie	dzienne
tak	tak	nie	mężczyzna	> 200 tys.	małopolskie	dzienne
tak	tak	tak	mężczyzna	> 200 tys.	małopolskie	dzienne
tak	nie	tak	kobieta	> 200 tys.	małopolskie	dzienne
tak	nie	tak	mężczyzna	do 50 tys.	mazowieckie	dzienne
tak	nie	tak	kobieta	> 200 tys.	śląskie	dzienne
nie	nie	tak	kobieta	do 50 tys.	podkarpackie	dzienne
tak	tak	tak	kobieta	> 200 tys.	małopolskie	dzienne
tak	tak	tak	kobieta	wieś	małopolskie	dzienne
tak	tak	tak	kobieta	do 50 tys.	świętokrzyskie	dzienne
tak	tak	tak	kobieta	wieś	śląskie	dzienne
tak	nie	tak	mężczyzna	> 200 tys.	małopolskie	dzienne
tak	nie	nie	kobieta	100-200 tys.	małopolskie	dzienne
nie	tak	nie	kobieta	> 200 tys.	małopolskie	dzienne

Model: CHAID

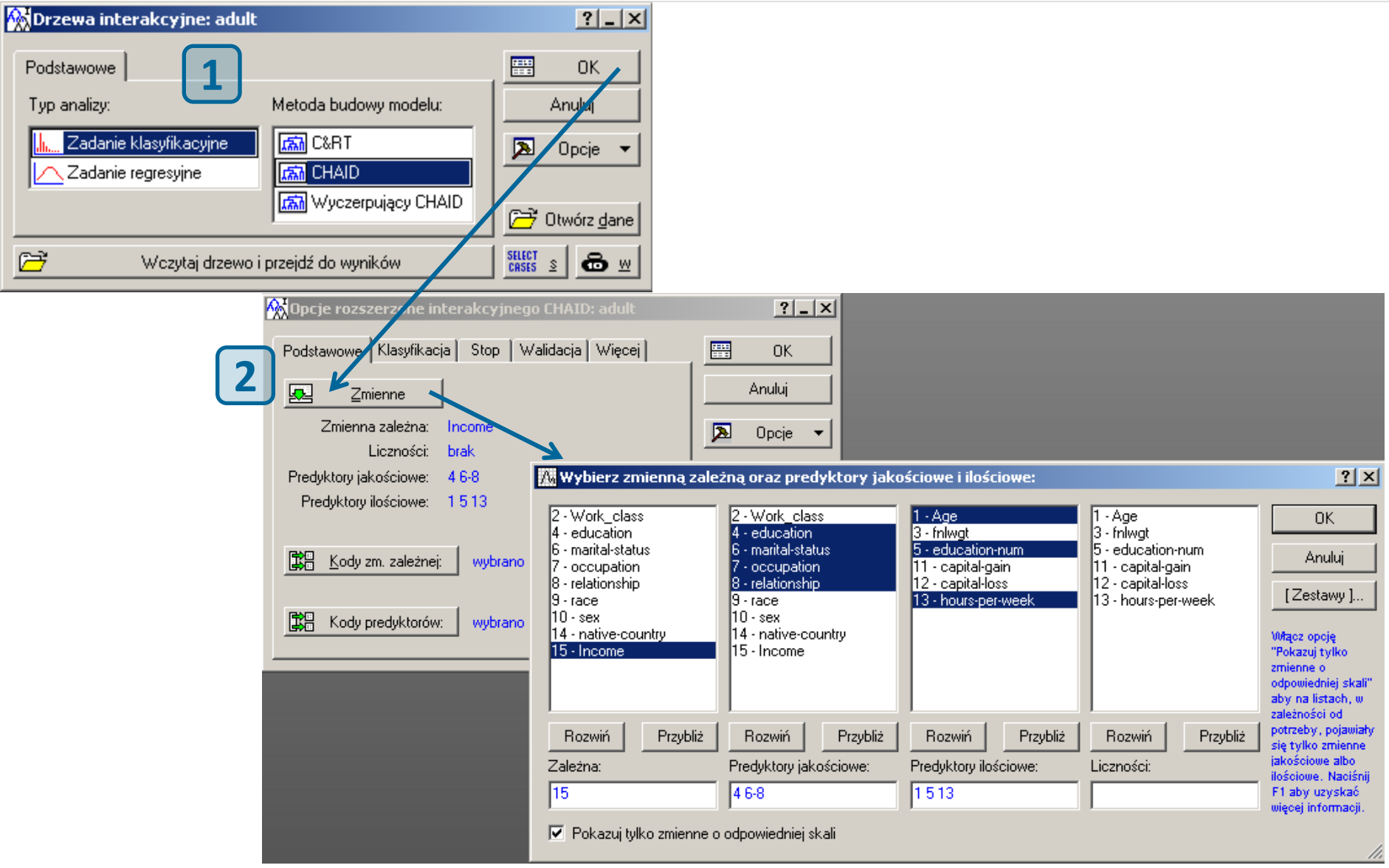
	Liczba węzły	Typ podziału	chi-kwadrat Statystyka	Stopnie Swobody	Skorygowane p
płeć	2	Automatycznie	27,13573	1,000000	0,000000
prawo jazdy	2	Automatycznie	21,86719	1,000000	0,000003
auto - badany	2	Automatycznie	17,78424	1,000000	0,000025
tryb studiów	3	Automatycznie	28,19740	2,000000	0,000001
województwo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

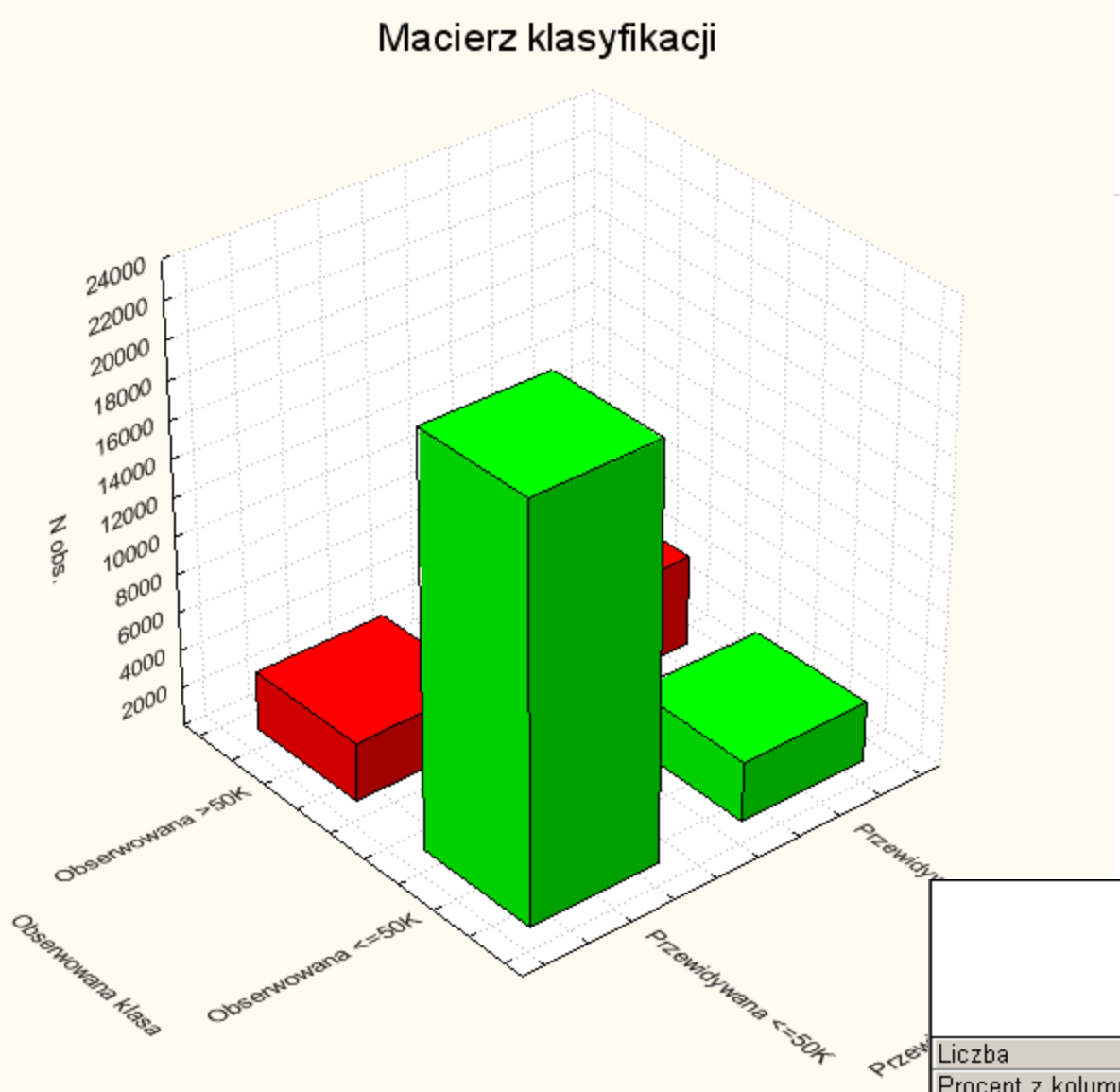




STATISTICA - inny przykład drzewa CHAID



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

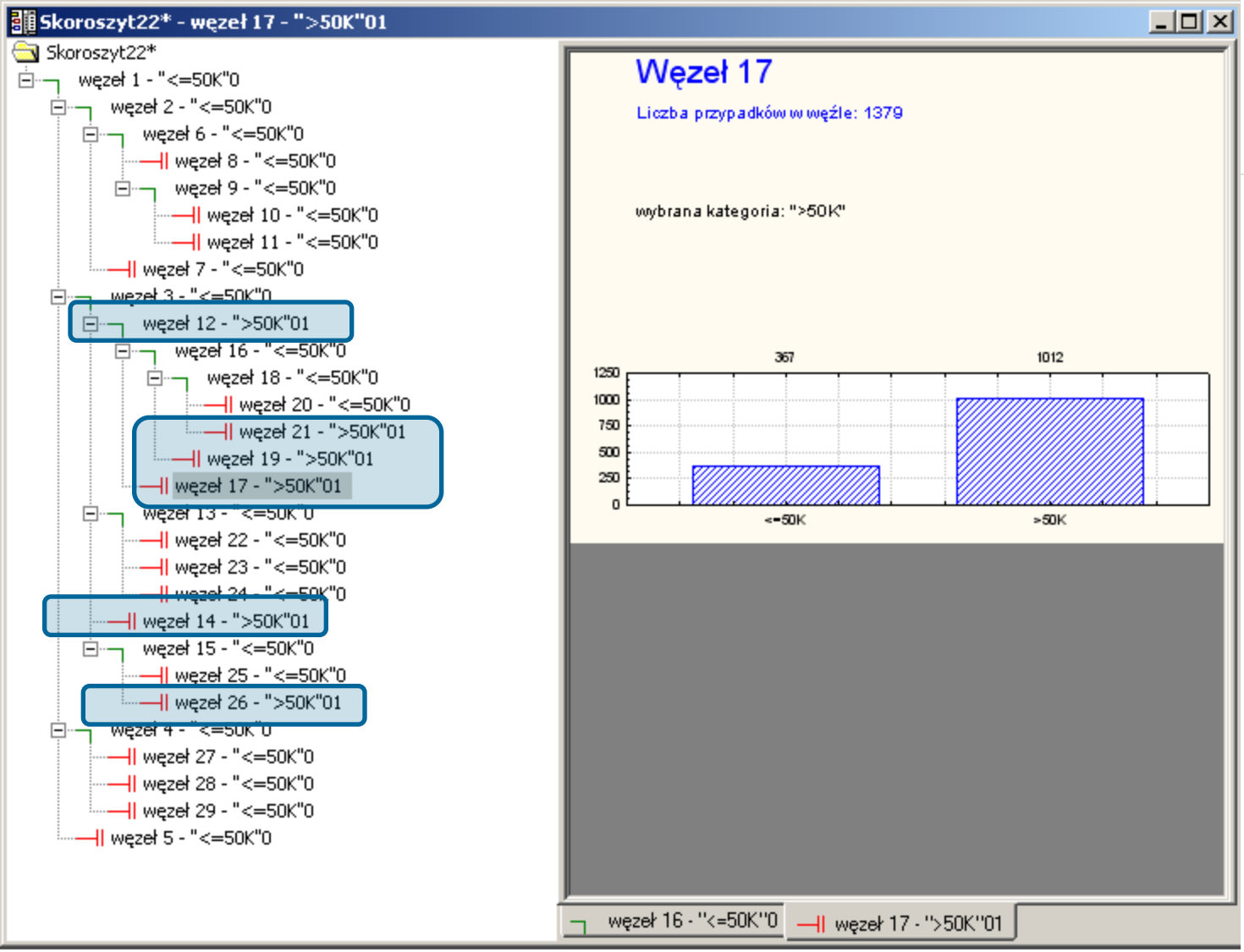


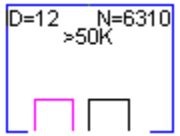
Macierz klasyfikacji (adult)
Zm. zal.:Income
Model: CHAID

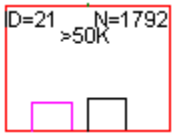
	Obserw.	Przewidywana <=50K	**Przewidywana >50K**	Łącznie w wierszu
Liczba	<=50K	21569	3151	24720
Procent z kolumnny		87.63%	39.65%	
Procent z wiersza		87.25%	12.75%	
Procent z ogółu		66.24%	9.68%	75.92%
Liczba	>50K	3045	4796	7841
Procent z kolumnny		12.37%	60.35%	
Procent z wiersza		38.83%	61.17%	
Procent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

FP 2

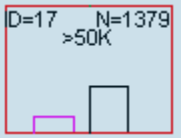
FN 1

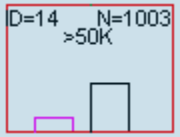

Skoroszyt22* - węzeł 17 - ">50K"01
Skoroszyt22*
węzeł 1 - "<=50K"0
węzeł 2 - "<=50K"0
węzeł 6 - "<=50K"0
węzeł 8 - "<=50K"0
węzeł 9 - "<=50K"0
węzeł 10 - "<=50K"0
węzeł 11 - "<=50K"0
węzeł 7 - "<=50K"0
węzeł 3 - "<=50K"0
węzeł 12 - ">50K"01
węzeł 16 - "<=50K"0
węzeł 18 - "<=50K"0
węzeł 20 - "<=50K"0
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

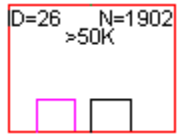

D=12 N=6310
>50K

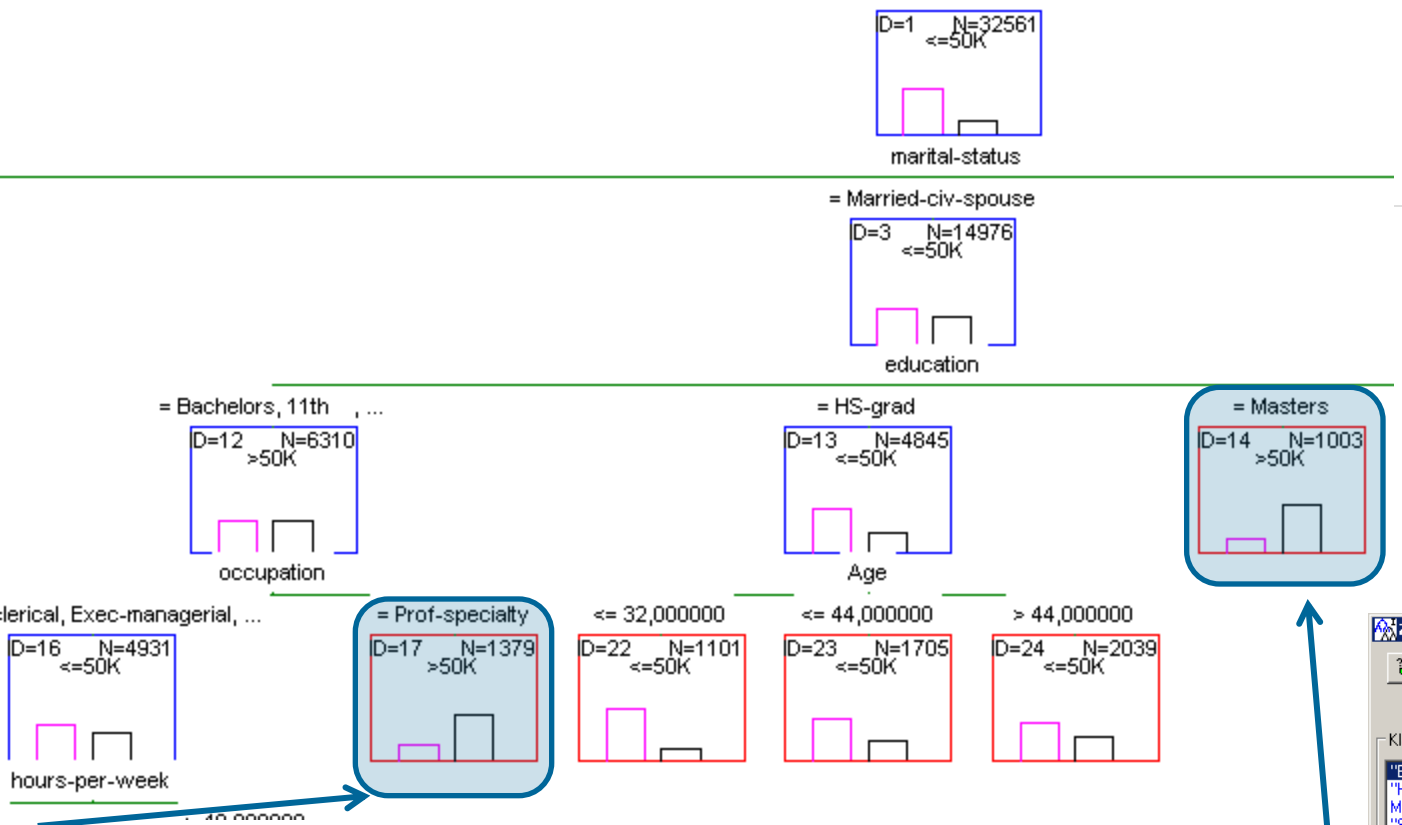

D=21 N=1792
>50K

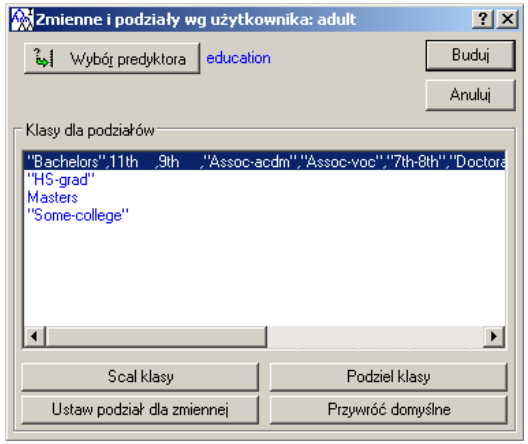

D=19 N=1871
>50K


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%)

Indukcja drzew regresyjnych

zmienna zależna: ilościowa

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.

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.

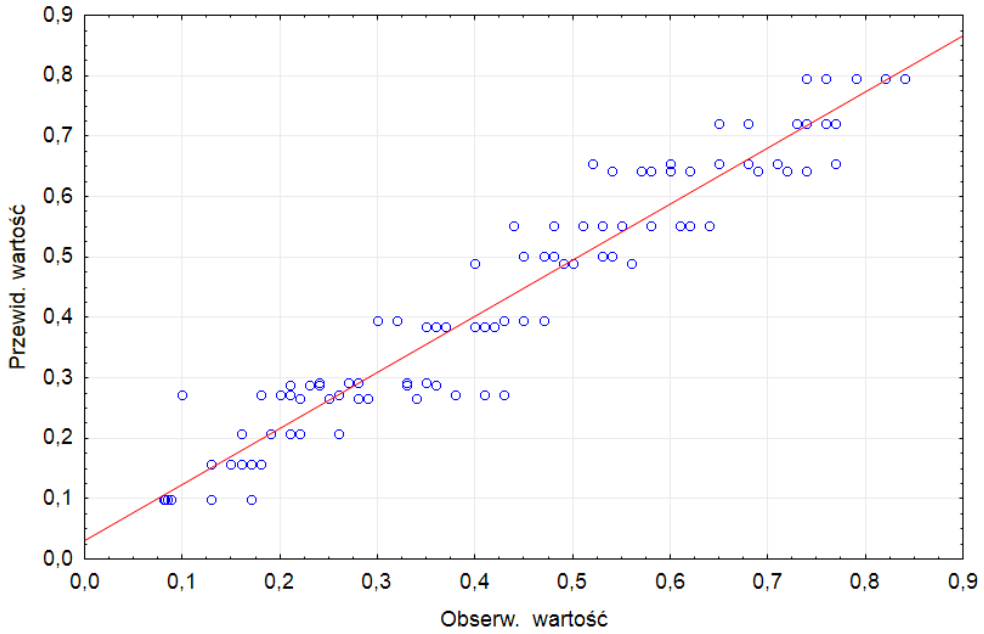
Współczynnik determinacji

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.

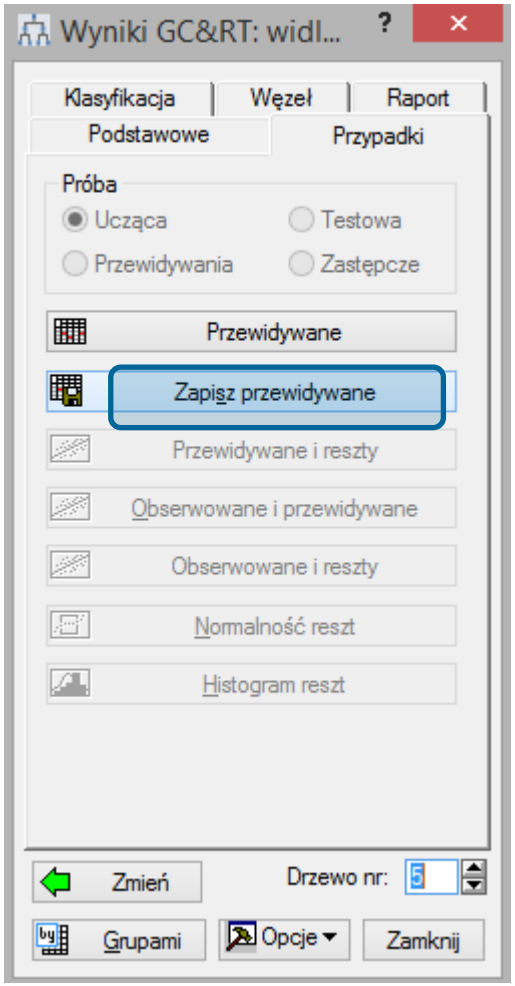
$$R^2 = \frac{\sum_{t=1}^{n}(\hat{y}_t - \overline{y})^2}{\sum_{t=1}^{n}(y_t - \overline{y})^2},$$

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

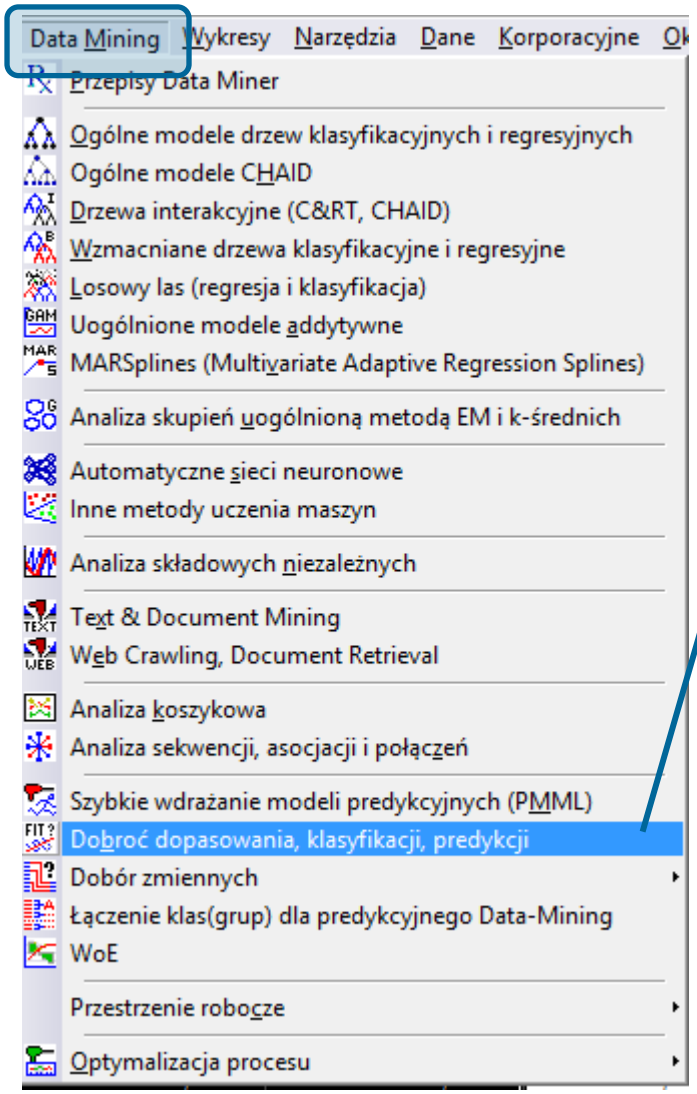


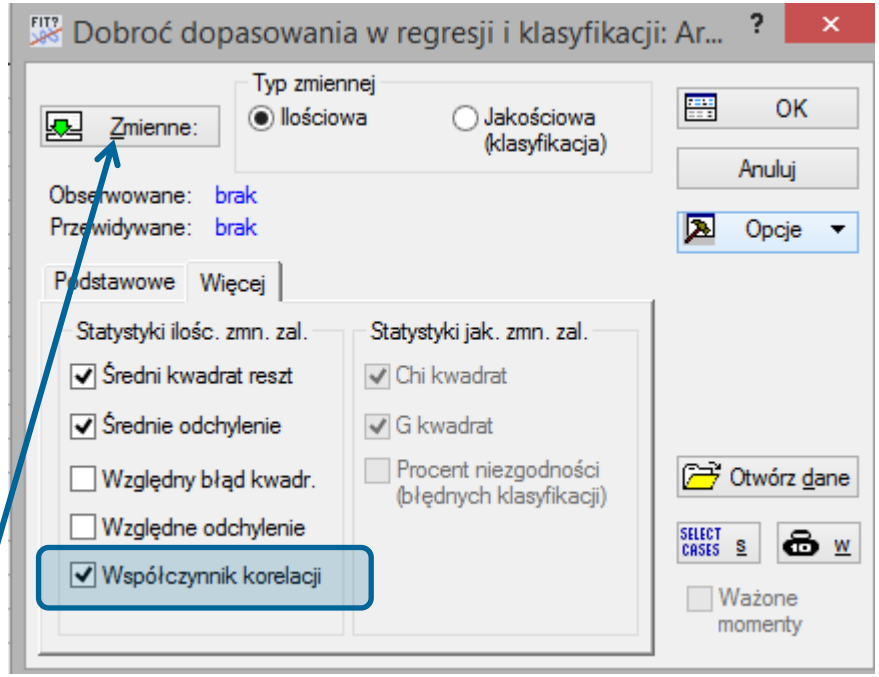
Dobroć dopasowania (1)



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)



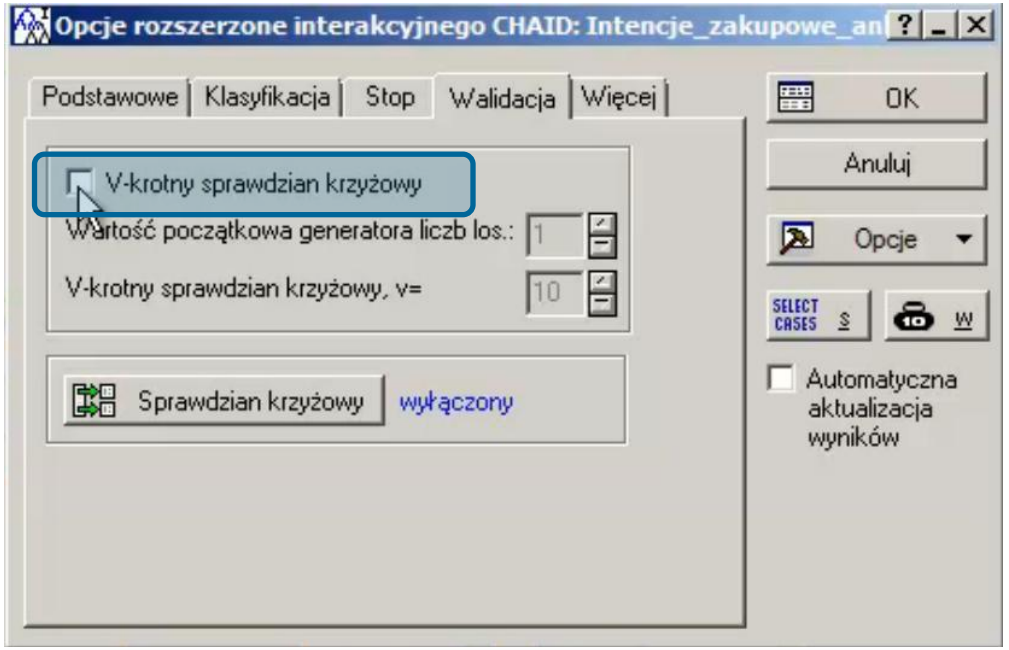


współczynnik determinacji
to kwadrat współczynnika korelacji

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.

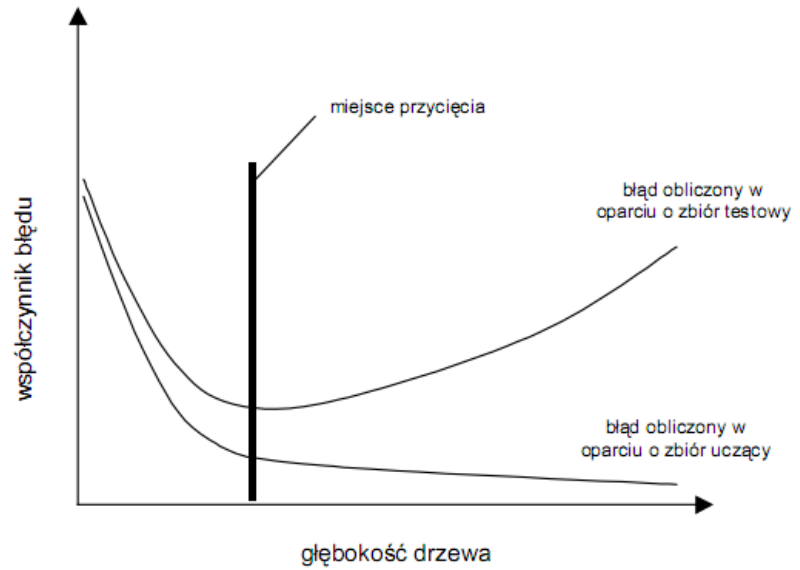
$$R(d) = \frac{1}{N}\sum_{i=1}^{N}(y_i - d(x_i))^2$$

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

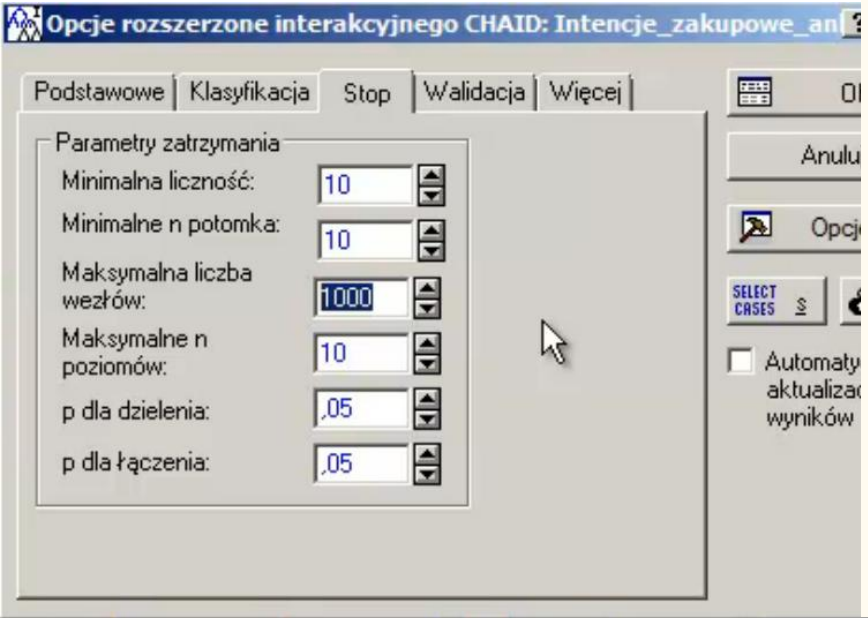
Miara ta obliczana jest dla tego samego zbioru danych, na bazie którego zbudowano model (partycję) d.

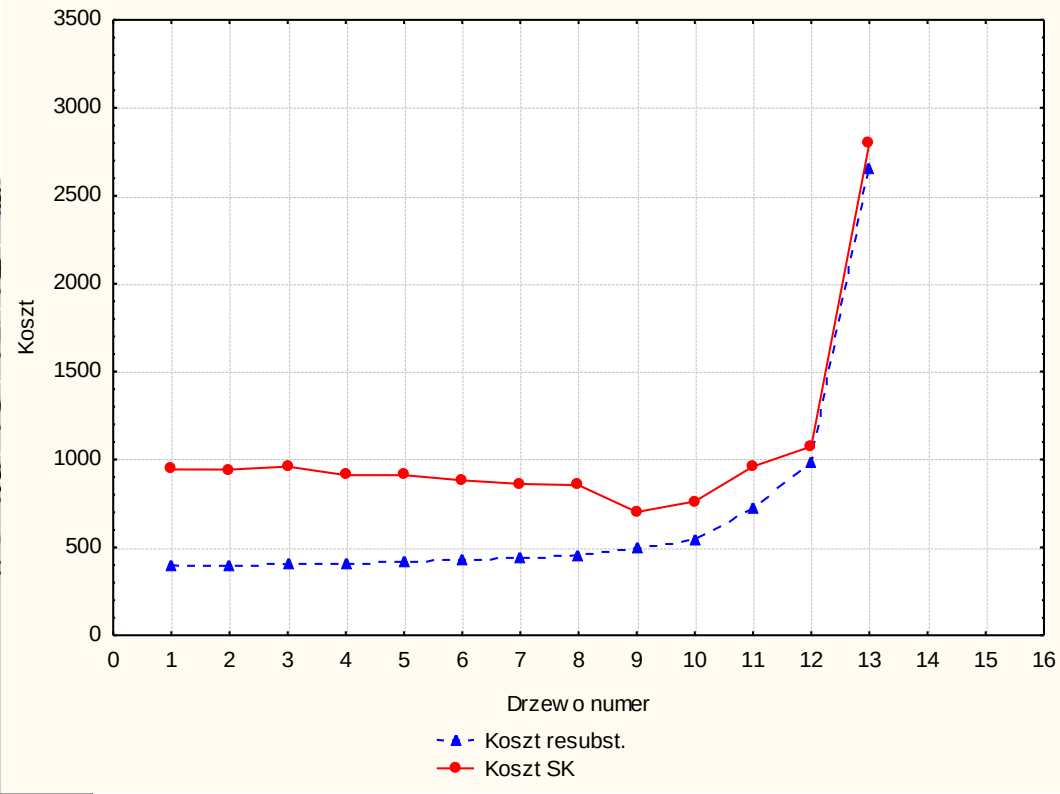
niski koszt resubstytucji = wartości zmiennej zależnej bliskie średniej w danym liściu

Wybór drzewa – przycinanie drzewa (1)

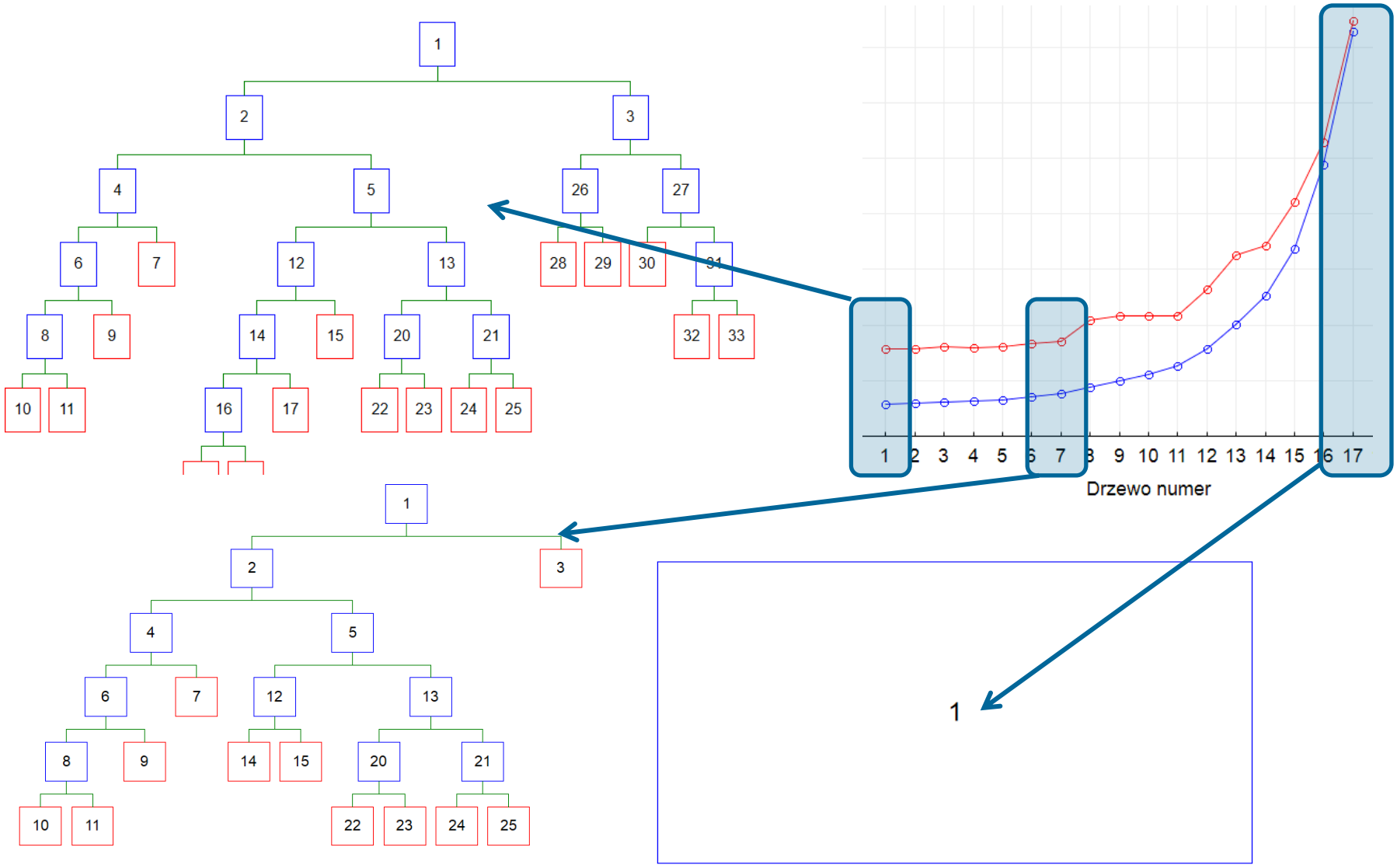


Jedną z metod doboru drzewa jest wybranie takiego, dla którego koszty resubstytucji i koszt sprawdzianu krzyżowego (SK) się przecinają.

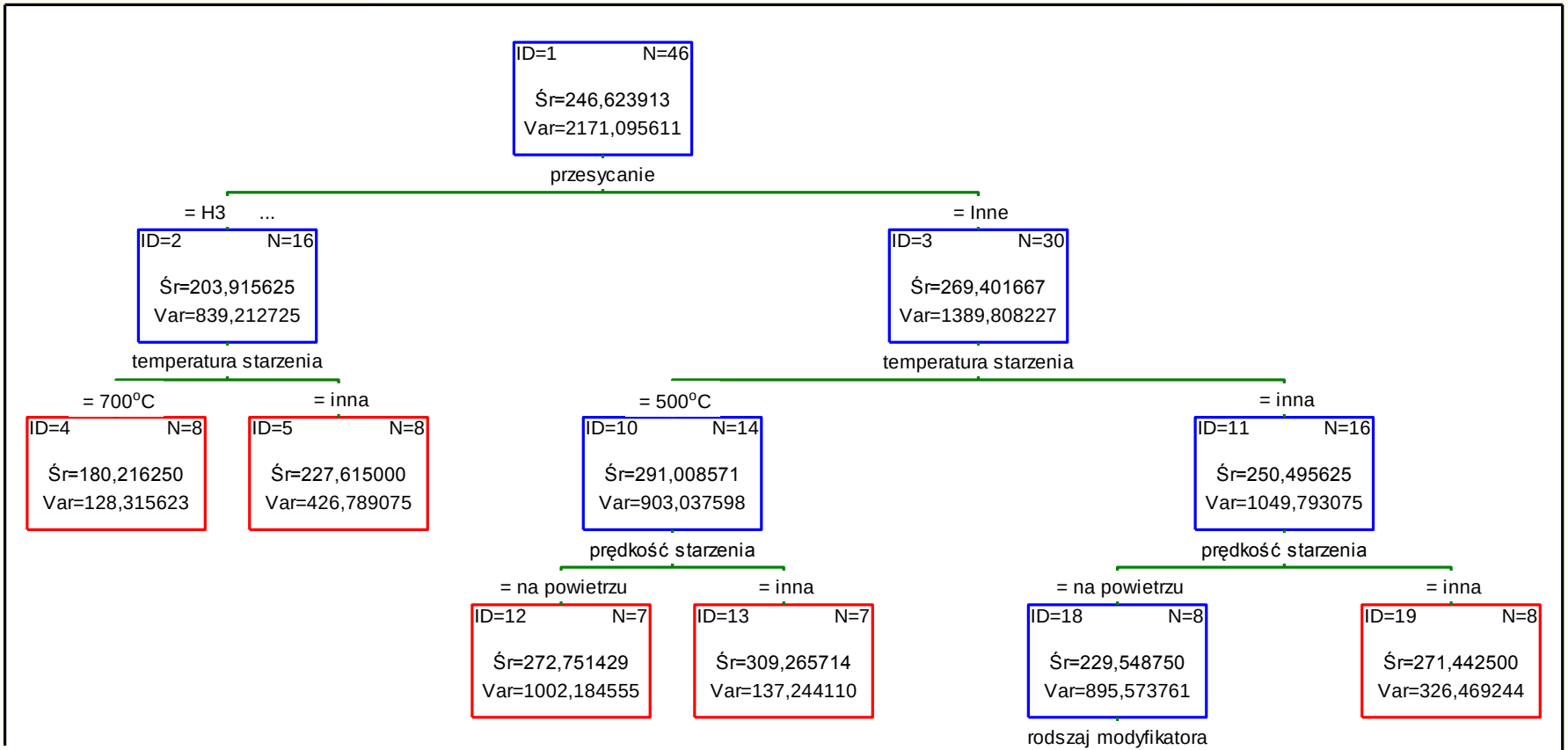




Wybór drzewa – przycinanie drzewa (2)



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

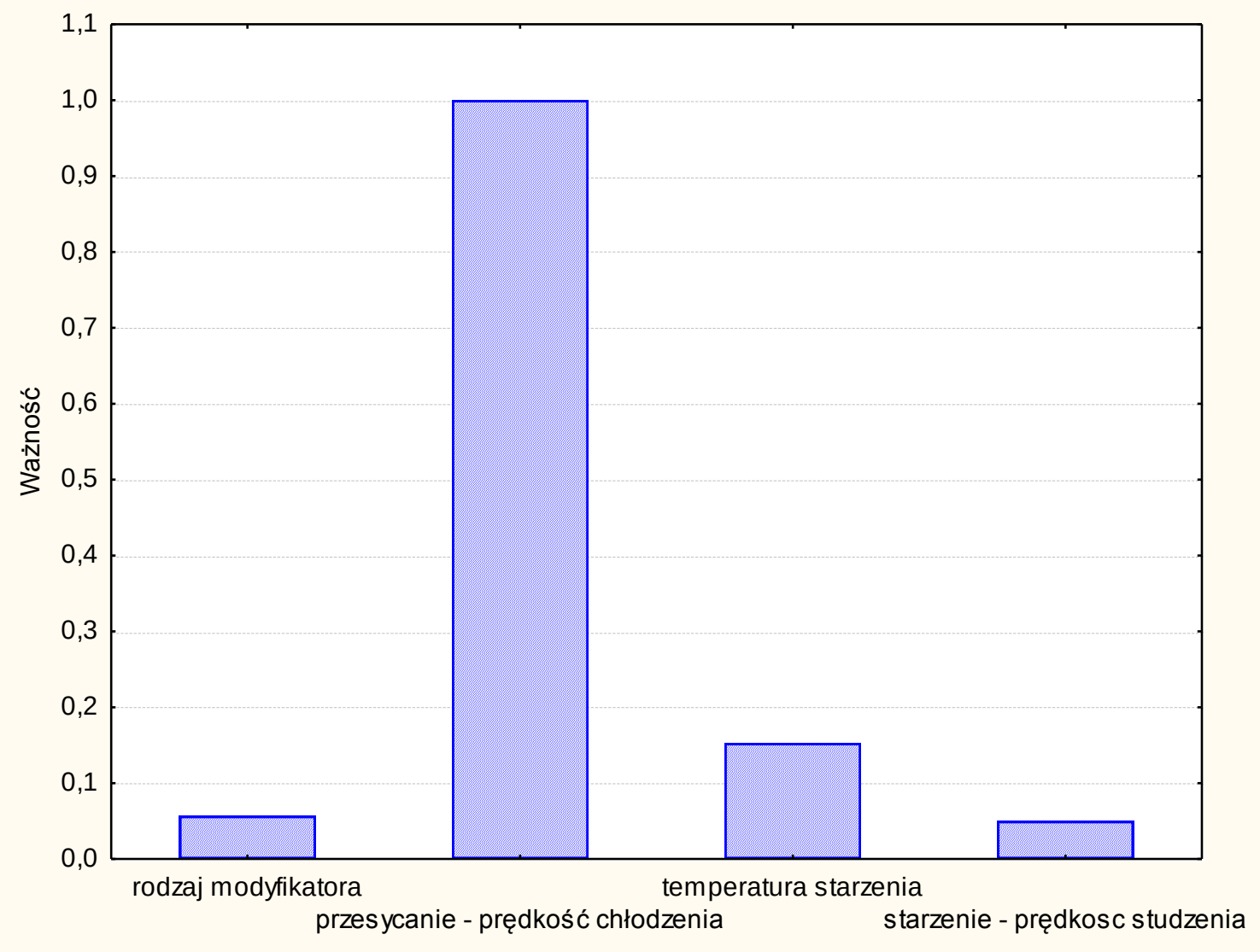


Klasy dla poszczególnych parametrów R_m, $R_{0,2}$, A zostały wyznaczone za pomocą modeli drzew regresyjnych w oparciu o zmienne predykcyjne jakimi były:

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

Co jeszcze? – Ważność predyktorów

- Algorytm drzewa C&RT pozwala określić ważność poszczególnych zmiennych predykcyjnych.
- Daną zmienną uznajemy za ważną w procesie klasyfikacji, czyli za niosącą informację o klasie, jeśli zmienna ta często bierze udział w procesie klasyfikowania obiektów ze zbioru uczącego.
- „Gotowość" atrybutu do brania udziału w procesie klasyfikacji mierzona jest w trakcie budowy drzew klasyfikacyjnych.
- Ważność oznacza wysoki stopień współzmienności (wyrażonej kowariancją lub korelacją) danego czynnika ze zmienną zależną, do ustalenia tego parametru służą takie techniki jak metody regresji wielorakiej czy algorytm względnej ważności Kruskala lub analiza dominacji.


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

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

univariate/multivariate

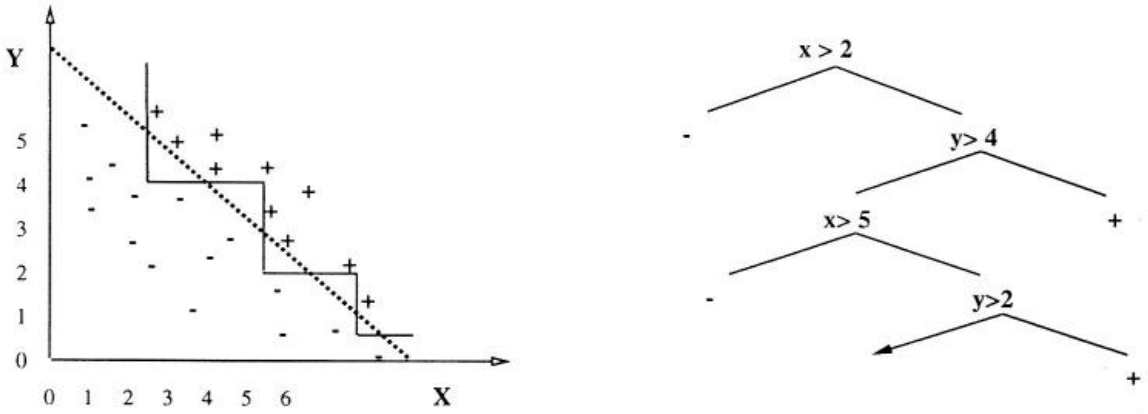
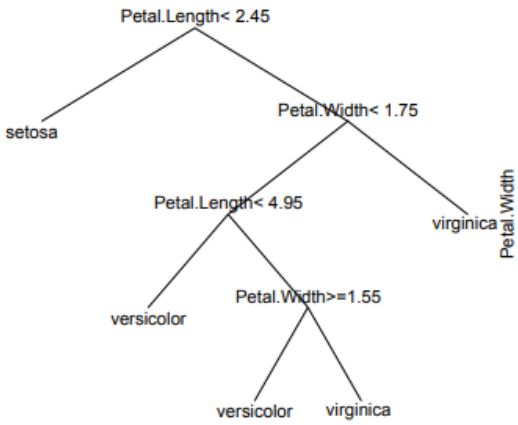
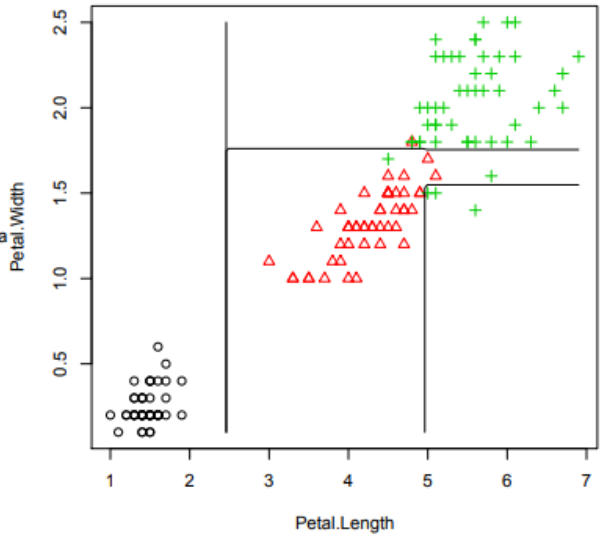


Figure 1. An example instance space; "+": positive instance, "−": negative instance and the corresponding univariate decision tree.

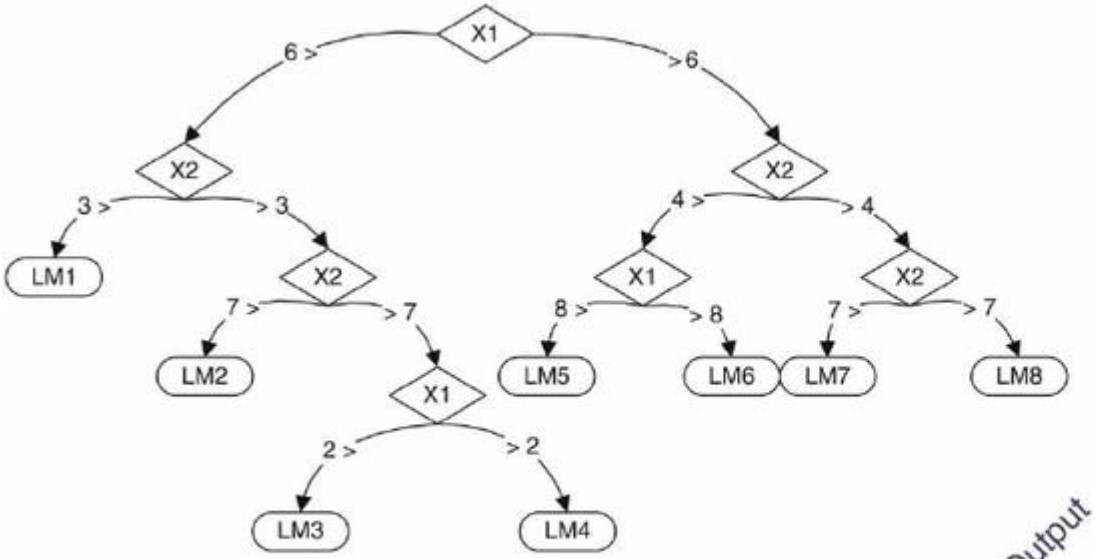
The multivariate decision tree-constructing algorithm selects not the best attribute but the best linear combination of the attributes.

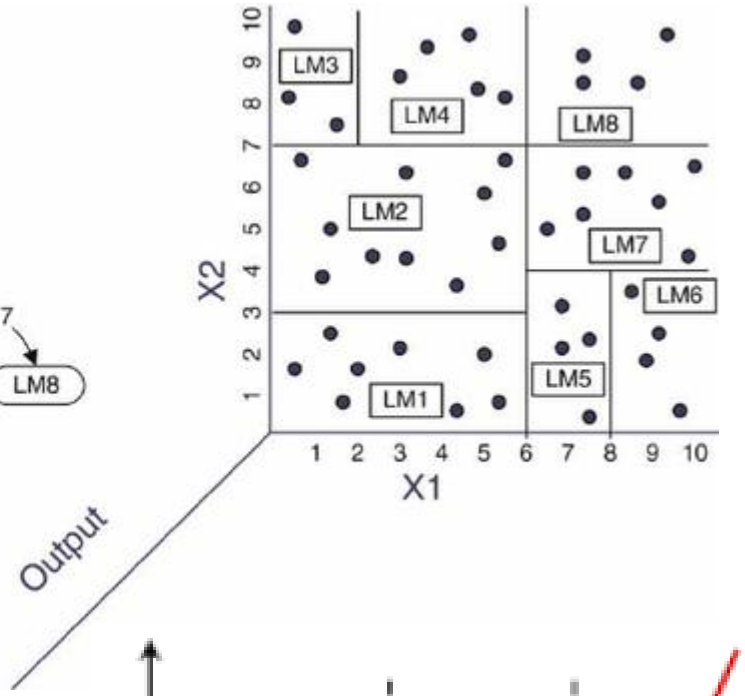




M5P Decision Tree

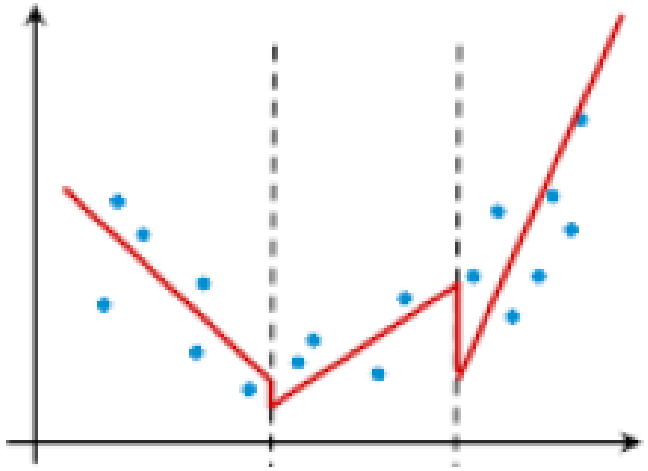
binary regression tree model





divergence metric:
Standard Deviation Reduction (SDR)

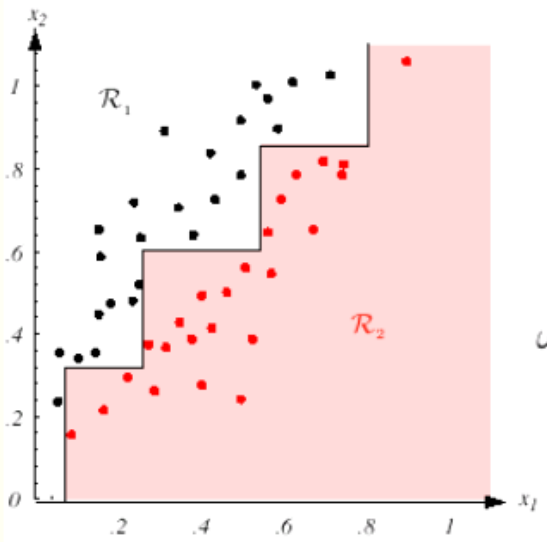
$$SDR = sd(T) - \frac{\sum_{i=1}^{n} |T_i|}{|T|} * sd(T)$$

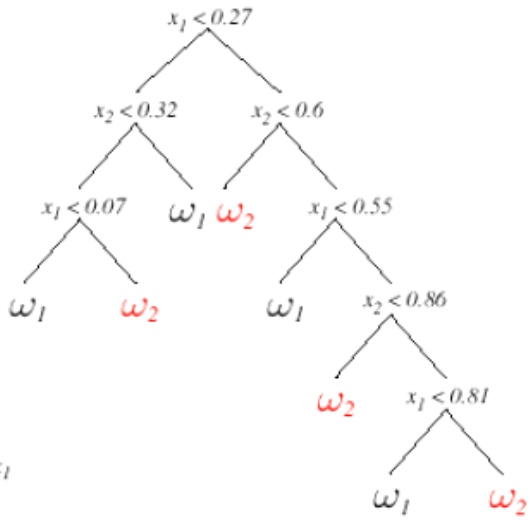


Oblique trees (skośne)

axis-parallel trees:

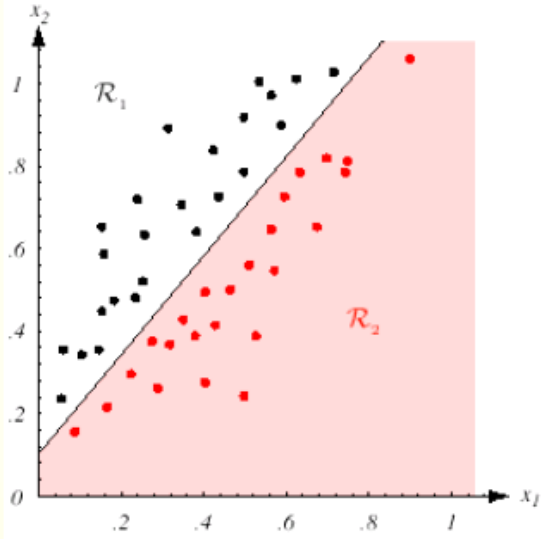
test na jednym atrybucie na raz

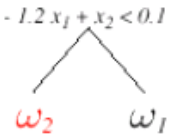


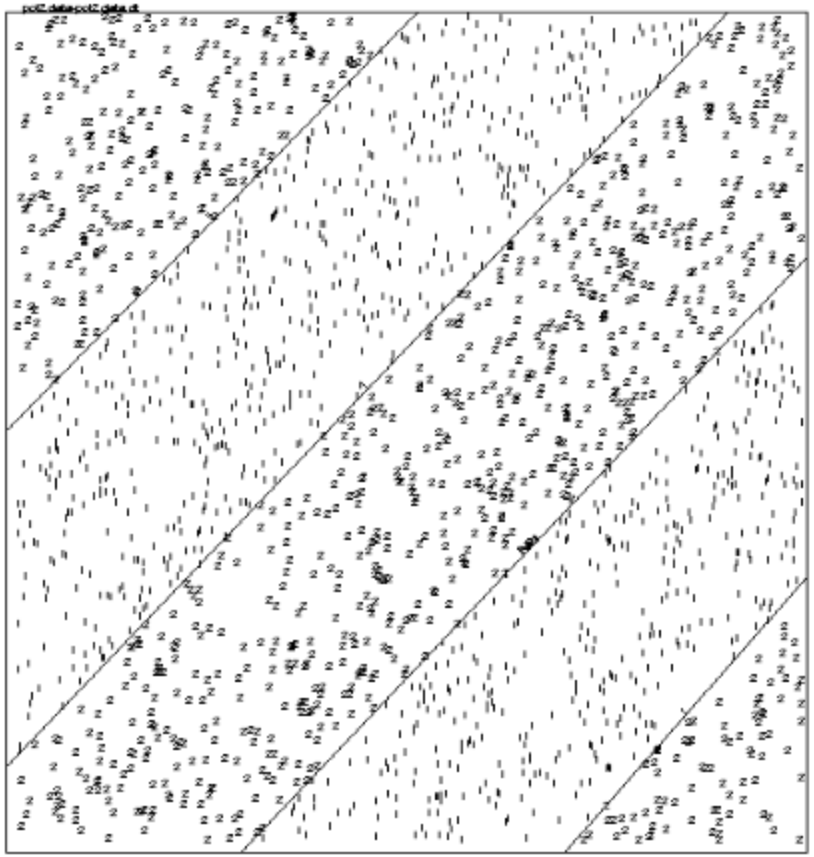


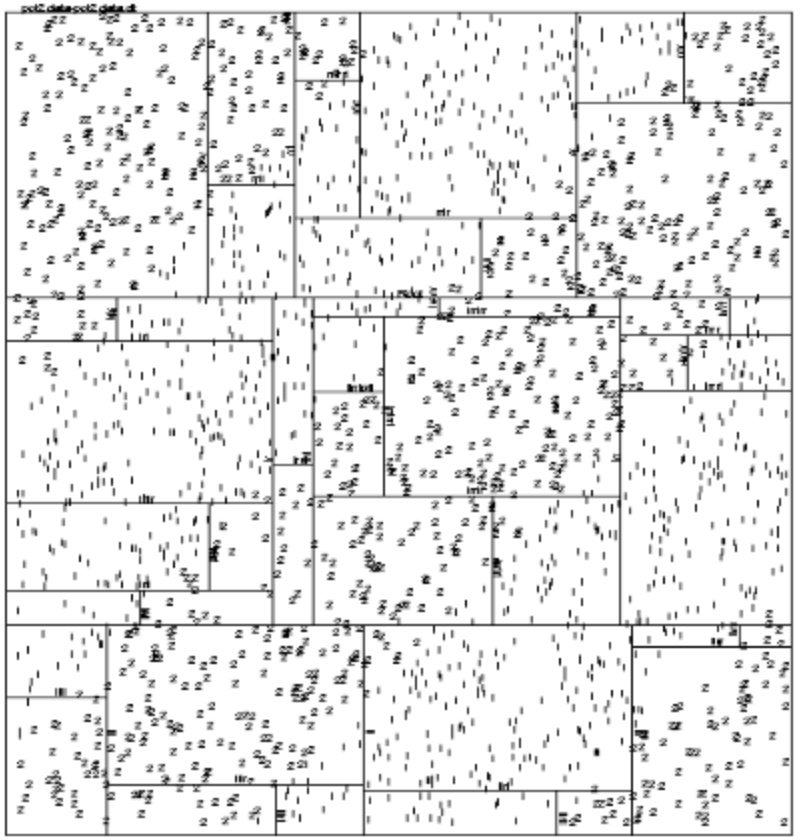
oblique trees:

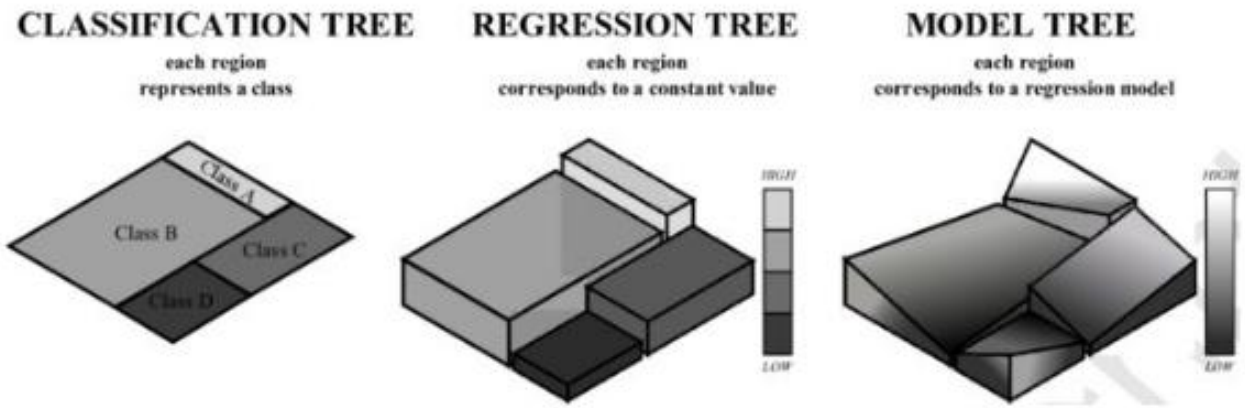
test jest funkcją (multivariate tests)

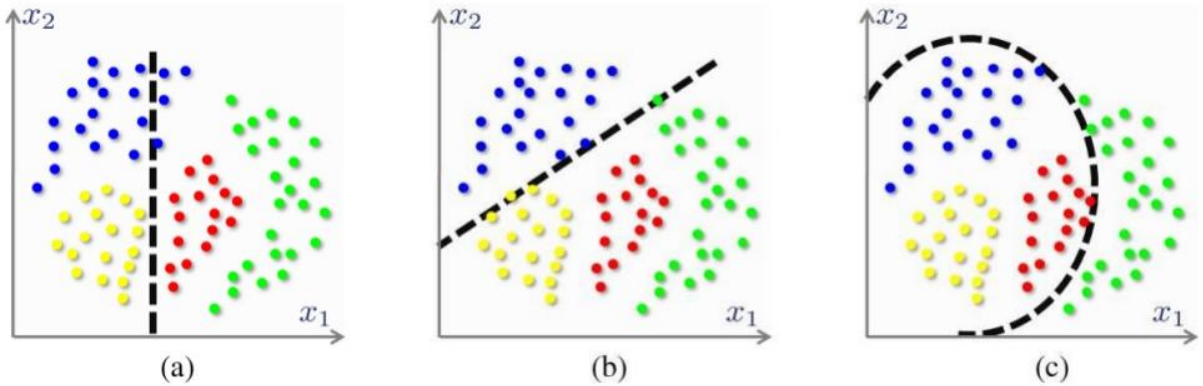









CLASSIFICATION TREE
each region
represents a class
Class A
Class B
Class C
Class D
REGRESSION TREE
each region
corresponds to a constant value
MODEL TREE
each region
corresponds to a regression model


x_2
x_1
(a)
Axis aligned split
x_2
x_1
(b)
Oblique split
x_2
x_1
(c)
Polynomial split

Discussion

TDIDT: Top-Down Induction of Decision Trees

The most extensively studied method of machine learning used in data mining

Different criteria for attribute/test selection rarely make a large difference

Different pruning methods mainly change the size of the resulting pruned tree

C4.5 (C5.0) builds *univariate* decision trees: each node tests a single attribute

Some TDITDT systems can build *multivariate* trees

(e.g., the famous CART tree learner, Oblique trees)

Central Quest

Find true patterns

and

avoid overfitting

(false patterns due to randomness)

Algorytmy indukcji reguł

- rule induction:
  - AQ – R. Michalski (1969)
  - CN2 — Clark, Niblett (1989)
  - RIPPER (IREP), RULE Extraction System,
  - PRISM (ID3), ACO (ants),
  - RISE, DeEPs, DeEPsNN, RIONA – unified approach
  - rough sets – Z. Pawlak (1982) (LEM, MODLEM, RSES, ROSETTA, etc.)
- instance based learning (IBL): kNN, IB3, PEBLS,
  - lazy rule induction approach
  - Bayesian learning
- indukcja drzew: CART, CHAID, C5.0, SLIQ, ID3, SPRINT, Oblique trees, Random Forest, Boosted Trees etc.
- multiple classifiers, multistrategy learning (combine approach):
  - ANFIS: adaptive neuro-fuzzy inference system,
  - ProbRough, MCS, ITRULE, KBNGE – empirical verification
- SNN (odzyskiwanie)
- regresja…

Uczenie

zagadnienia: reprezentacja, przeszukiwanie, walidacja

reprezentacja: decision trees, sets of rules, instances, neural networks

przeszukiwanie: learning algorithm finds the concept description in a space of possible descriptions defined by the representation language

walidacja: miary jakości kandydatów

Indukcja reguł

indukcja reguł – małe zbiory silnych predyktorów

podziały równoległe do osi (np. $x_n>5$)

problem z obserwacjami odstającymi i mało licznymi sekcjami

odporne na szum

liczba podziałów rośnie szybkim tempie

właściwe podejście dla danych jakościowych (symbolic)

i zmiennych o małej istotności

IBL – instance based learning

klasyfikatory minimalnoodległościowe (np. kNN)

dobrze radzą sobie z nieliniowością i „outliners”

wrażliwe na szum

wrażliwe na zmienne nieistotne

dobre dla danych numeryczny

complete set of consistent and minimal decision rule

size of the minimal rules set can be exponential with respect to the size of the training set - rule set that is not necessarily complete

memory based (lazy concept induction) - do not require calculation of the decision rule set before classification of new objects

generates only decision rules relevant for a new test object and then classifies it like algorithms generating rules in advance

Instance Based Learning (IBL):

lazy rule induction approach (kNN)

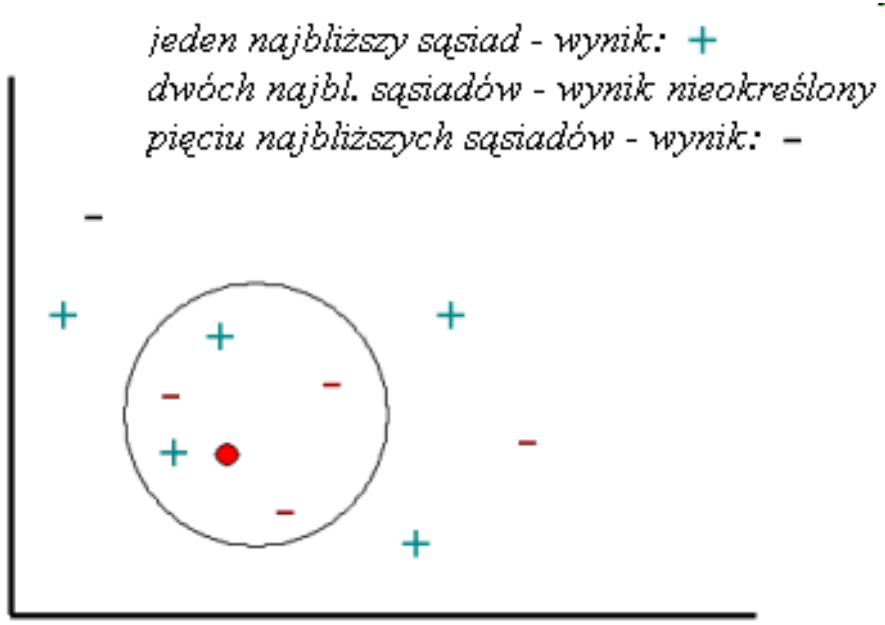
Bayesian learning

Klasyfikatory kNN

1. Klasyfikator *kNN - klasyfikator k-najbliższych sąsiadów* (ang. *k-nearest neighbor classifier)*
2. Klasyfikacja nowych przypadków jest realizowana „na bieżąco", tj. wtedy, gdy pojawia się potrzeba klasyfikacji nowego przypadku.
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.

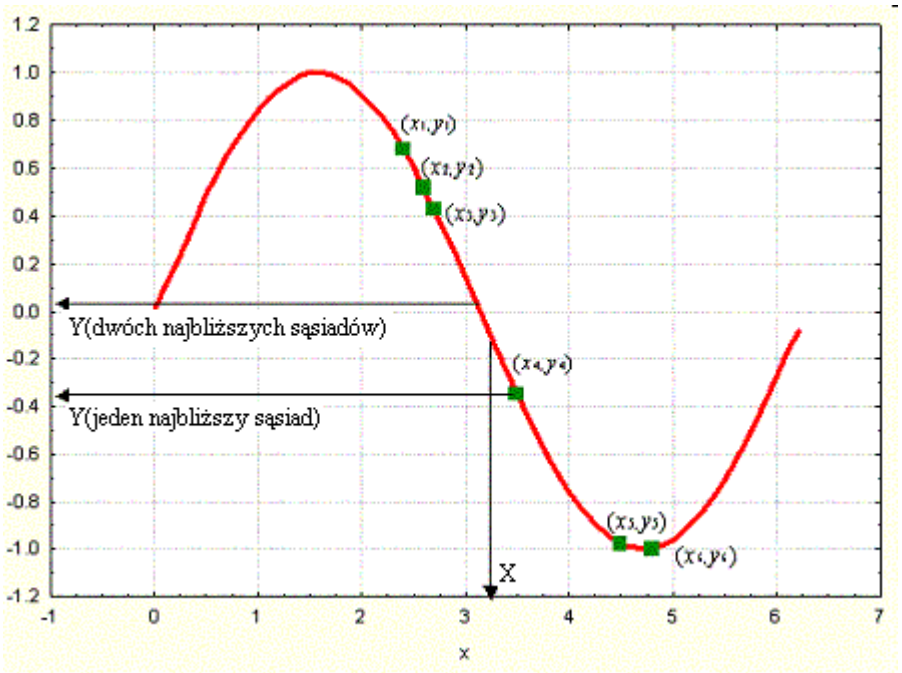


1. Zacznijmy od rozpatrzenia przypadku jednego, najbliższego sąsiada. Widać, że najbliżej czerwonego punktu jest plus, tak więc nowy przypadek zostanie zaklasyfikowany do plusów.

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 sposób postępujemy 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.

Problemy związane z klasyfikatorem kNN:

» jak zdefiniować punkt „najbliższy" nowemu przykładowi X?

» problemem transformacji: Jak przetransformować przykład do punktu w przestrzeni wzorców?

definicja funkcji odległości:

– klasyfikatory kNN stosują najczęściej euklidesową miarę odległości, czyli po prostu odległość geometryczna w przestrzeni wielowymiarowej.

$$odległość(x,y) = \{\Sigma_i (x_i - y_i)^2\}^{1/2}$$

– Odległość euklidesową podnosi się do kwadratu, aby przypisać większą wagę obiektom, które są bardziej oddalone

Miary odległości - Przestrzenie metryczne

- Odległość elementów przestrzeni cech ilościowych
  - » odległość euklidesowa
  - » odległości Sebestyena
  - » odległość uogólnioną (Mahalanobisa)
  - » odległość Manhattan
  - » odległość Canbera
  - » Odległość Chernoffa
  - » Odległość Bhattacharyya
  - » Dywergencja Kullbacka-Leiblera (zwana też entropią względną lub relatywną entropią)

- Odległość elementów przestrzeni cech jakościowych
  - » odległości Hamminga
  - » SVDM metric
  - » Odległość cosinusowa
  - » Odległość Levenshteina (edycyjna) – miara odmienności napisów (skończonych ciągów znaków)odległości Sebestyena
  - » Odległość Damerau-Levenshteina
  - » Niezgodność procentowa
  - » Chi^2
  - » miara VDM (Value Difference Metric).
  - » Miara Lance'a i Williamsa

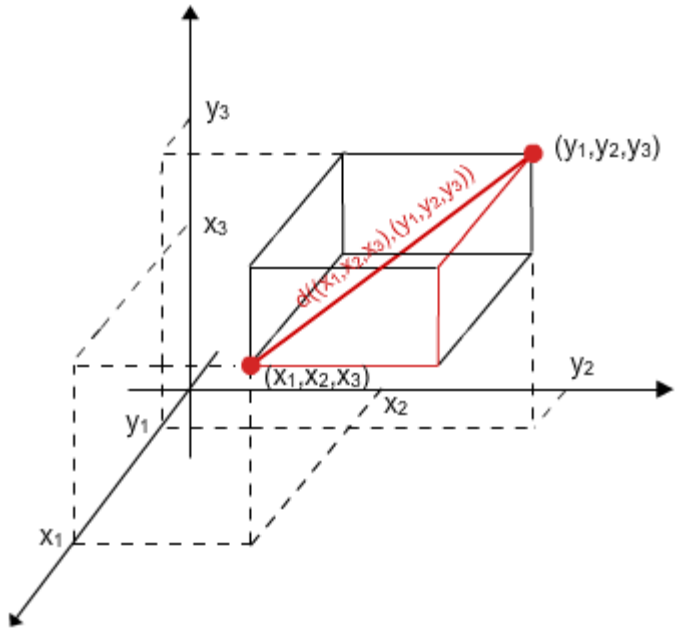
Miary odległości - Przestrzenie metryczne

Odległość miejska (Manhattan, City block). Ta odległość jest sumą różnic mierzonych wzdłuż wymiarów.

W większości przypadków ta miara odległości daje podobne wyniki, jak zwykła odległość euklidesowa.

w przypadku tej miary, wpływ pojedynczych dużych różnic (przypadków odstających) jest stłumiony (ponieważ nie podnosi się ich do kwadratu).

$$odległość(x,y) = \Sigma_i \, |x_i - y_i|$$



Standaryzacja / Normalizacja

W wyniku **normalizacji** danych otrzymujemy wektory, których wartości cech są zawarte w **przedziale <0,1>**. Normalizacja nie uwzględnia rozkładu wartości danej cechy.

$$u_i = \frac{x_i}{x_{\max} - x_{\min}} \qquad a'_i = \frac{a_i - a_{i\_\min}}{a_{i\_\max} - a_{i\_\min}} \qquad u_i = \frac{x_i}{\sum_{i=1}^{n} x_i}$$

Wynikiem **standaryzacji** jest wektor cech , których wartość średnia ***m = 0*** , natomiast odchylenie standardowe ***s = 1***, dzięki czemu wszystkie cechy mają jednakowy wkład do wartości odległości

$$z_i = \frac{x_i - \bar{x}}{S(x)}$$

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

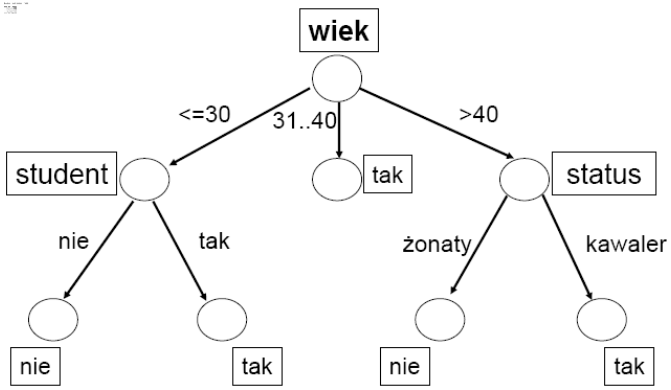
Naiwny klasyfikator Bayesa

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



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 $i=1,2$

Przykład

wiek	dochód	student	status	kupi_komputer
<=30	wysoki	nie	kawaler	nie
<=30	wysoki	nie	żonaty	nie
31..40	wysoki	nie	kawaler	tak
>40	średni	nie	kawaler	tak
>40	niski	tak	kawaler	tak
>40	niski	tak	żonaty	nie
31..40	niski	tak	żonaty	tak
<=30	średni	nie	kawaler	nie
<=30	niski	tak	kawaler	tak
>40	średni	tak	kawaler	tak
<=30	średni	tak	żonaty	tak
31..40	średni	nie	żonaty	tak
31..40	wysoki	tak	kawaler	tak
>40	średni	nie	żonaty	nie

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’

Probabilities for weather data

Outlook			Temperature			Humidity			Windy			Play	
	Yes	*No*		*Yes*	*No*		*Yes*	*No*		*Yes*	*No*	*Yes*	*No*
Sunny	2	3	Hot	2	2	High	3	4	False	6	2	9	5
Overcast	4	0	Mild	4	2	Normal	6	1	True	3	3		
Rainy	3	2	Cool	3	1								
Sunny	2/9	3/5	Hot	2/9	2/5	High	3/9	4/5	False	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	Mild	4/9	2/5	Normal	6/9	1/5	True	3/9	3/5		
Rainy	3/9	2/5	Cool	3/9	1/5								

Outlook	Temp	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

Probabilities for weather data

Outlook			Temperature			Humidity			Windy			Play	
	Yes	*No*		*Yes*	*No*		*Yes*	*No*		*Yes*	*No*	*Yes*	*No*
Sunny	2	3	Hot	2	2	High	3	4	False	6	2	9	5
Overcast	4	0	Mild	4	2	Normal	6	1	True	3	3		
Rainy	3	2	Cool	3	1								
Sunny	2/9	3/5	Hot	2/9	2/5	High	3/9	4/5	False	6/9	2/5	9/	5/
Overcast	4/9	0/5	Mild	4/9	2/5	Normal	6/9	1/5	True	3/9	3/5	14	14
Rainy	3/9	2/5	Cool	3/9	1/5								

- A new day:

Outlook	Temp.	Humidity	Windy	Play
Sunny	Cool	High	True	?

Likelihood of the two classes

For “yes” = $2/9 \times 3/9 \times 3/9 \times 3/9 \times 9/14 = 0.0053$

For “no” = $3/5 \times 1/5 \times 4/5 \times 3/5 \times 5/14 = 0.0206$

Conversion into a probability by normalization:

P(“yes”) = 0.0053 / (0.0053 + 0.0206) = 0.205

P(“no”) = 0.0206 / (0.0053 + 0.0206) = 0.795

102

Twierdzenie Bayesa: $P(C|X)=[P(X|C)\times P(C)]/P(X)$

$P(C)$ – prawdopodobieństwo *a priori*

$P(C|X)$ – prawdopodobieństwo *a posteriori* (gdy wiemy, że zdarzyło się X)

Frequency Table		Play Golf	
		Yes	No
Outlook	Sunny	3	2
	Overcast	4	0
	Rainy	2	3

$$P(x \mid c) = P(Sunny \mid Yes) = 3/9 = 0.33$$

Likelihood Table		Play Golf		
		Yes	No	
Outlook	Sunny	3/9	2/5	5/14
	Overcast	4/9	0/5	4/14
	Rainy	2/9	3/5	5/14
		9/14	5/14	

$$P(x) = P(Sunny) = 5/14 = 0.36$$

$$P(c) = P(Yes) = 9/14 = 0.64$$

Posterior Probability: $P(c \mid x) = P(Yes \mid Sunny) = 0.33 \times 0.64 \div 0.36 = 0.60$

„zero-frequency problem”

Twierdzenie Bayesa: $P(C|X)=[P(X|C)\times P(C)]/P(X)$
$P(C)$ – prawdopodobieństwo *a priori*
$P(C|X)$ – prawdopodobieństwo *a posteriori* (gdy wiemy, że zdarzyło się X)

- What if an attribute value does not occur with every class value? (e.g., “Outlook=overcast” for class “no”)
  - Probability will be zero: $P(\text{Outlook=Overcast}|no)=0$
  - *A posteriori* probability will also be zero: $P(no|X)=0$
    (Regardless of how likely the other values are!)
- Remedy: add 1 to the count for every attribute value-class combination (*Laplacian smoothing*)
- Result: probabilities will never be zero
- Additional advantage: stabilizes probability estimates computed from small samples of data

Information Gain – przykład (1)

zachmurzenie	słonecznie	1,2,8,9,11	3 N + 2 T	5/14
	pochmurno	3,7,12,13	4 T + 0 N	4/14
	deszczowo	4,5,6,10,14	3 T + 2 N	5/14
temperatura	gorąco	1,2,3,13	2 N + 2 T	4/14
	łagodnie	4,8,10,11,12,14	4 T + 2 N	6/14
	zimno	5,6,7,9	3 T + 1 N	4/14
wilgotność	wysoka	1,2,3,4,8,12,14	3 N + 4 T	7/14
	normalna	5,6,7,9,10,11,13	6 T + 1 N	7/14
wiatr	słaby	1,3,4,5,8,9,10,13	2 N + 6 T	8/14
	silny	2,6,7,11,12,14	3 T + 3 N	6/14

Entropia (rozkład):

$$Ent(S) = -\sum_{i=1}^{k} p_i \lg_2 p_i,$$

Information Gain – przykład (2)

W przykładzie *golf* jako pierwszy do podziału został wybrany atrybut *„zachmurzenie”*, bo jego wskaźnik *„gain”* był największy:
S – zawiera 14 elementów; 2 klasy – TAK (9 elementów) i NIE (5 elementów)

E(S) = -9/14 log 9/14 – 5/14 log 5/14 = 0.94

E(S/zachmurzenie) = 5/14(-3/5log23/5 – 2/5log22/5) + 4/14(-1log21 – 0log 20)
+ 5/14(-3/5log23/5 – 2/5log22/5) = 0.2

E(S/temperatura) = 4/14(-2/4log22/4 – 2/4log22/4) + 4/14(-3/4log23/4 – 1/4 log21/4)
+ 6/14(-2/6log22/6 – 4/6log24/6) = 0.48

E(S/wilgotnosc) = 7/14(-4/7log24/7 – 3/7log23/7) + 7/14(-6/7log26/7 – 1/7 log21/7) = 0.43

E(S/wiatr) = 8/14(-6/8log26/8 – 2/8log22/8) + 6/14(-3/6log23/6 – 3/6log2 3/6) = 0.71

Gain Information(zachmurzenie) = 0.94 – 0.2 = 0.74

Gain Information(temperatura) = 0.94 – 0.48 = 0.46

Gain Information(wilgotnosc) = 0.94 – 0.43 = 0.51

Gain Information(wiatr) = 0.94 – 0.71 = 0.23

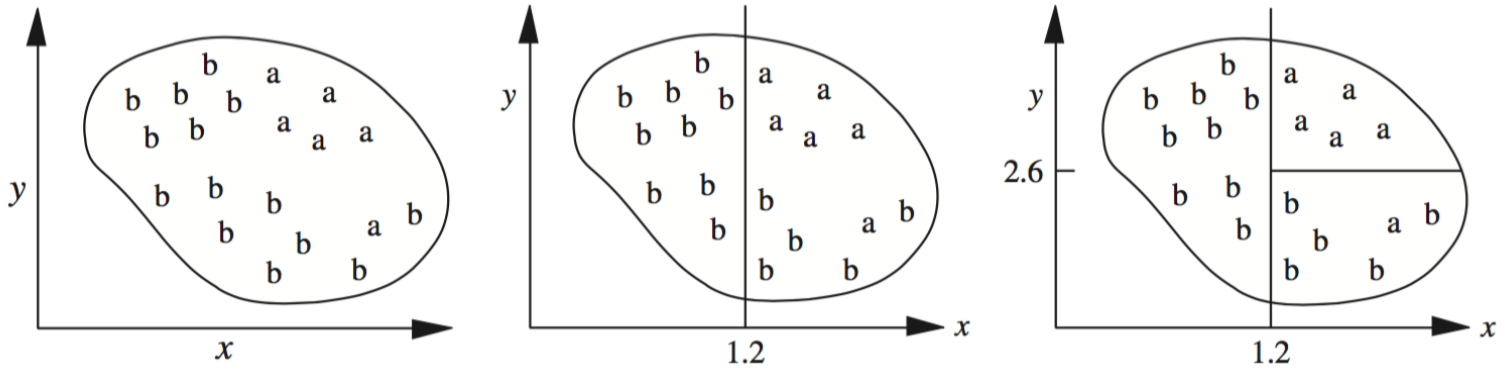
Największy zysk informacji dostarcza atrybut „zachmurzenie” i to on będzie korzeniem drzewa…

Algorytmy sekwencyjnego pokrywania

Covering algorithms

Sekwencyjne pokrywanie
Covering algorithms

- Instead, we can generate rule set directly
  - One approach: for each class in turn, find rule set that covers all instances in it
    (excluding instances not in the class)
- Called a *covering* approach:
  - At each stage of the algorithm, a rule is identified that “covers” some of the instances



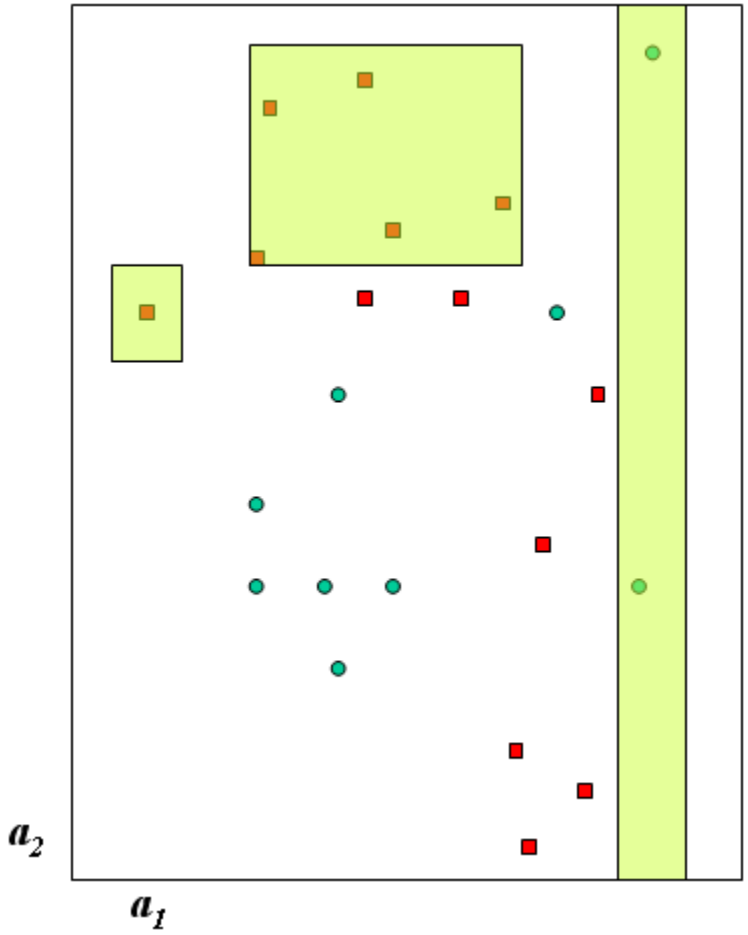

```
If x > 1.2 and y > 2.6
then class = a
```

przybliżone rozwiązywanie problemu pokrycia

Ryszard Michalski: założyciel i wieloletni dyrektor Laboratorium Uczenia Maszynowego na Uniwersytecie im. George'a Masona w USA, współpracownik zagraniczny Instytutu Informatyki PAN.

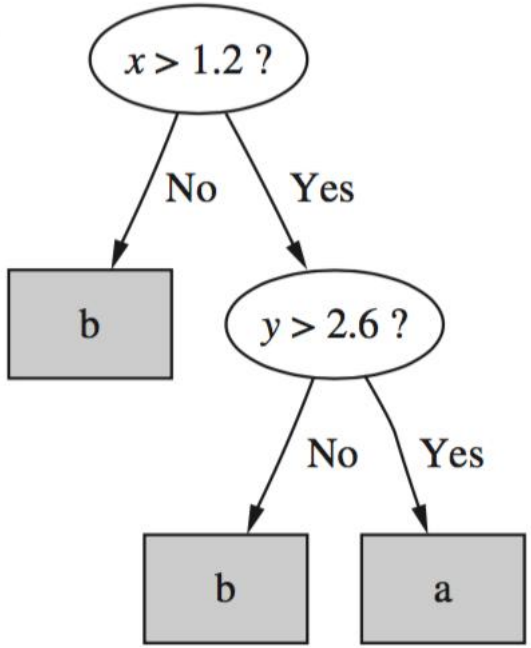
Algorytm AQ: (1969) oparty jest na pokrywaniu sekwencyjnym

for each class Ki **do**
    Ei := Pi U Ni (Pi positive, Ni negative example)
    RuleSet(Ki) := empty
    repeat {find-set-of-rules}
        find-one-rule R covering some positive examples
        and no negative ones
        add R to RuleSet(Ki)
        delete from Pi all pos. ex. covered by R
    until Pi (set of pos. ex.) = empty



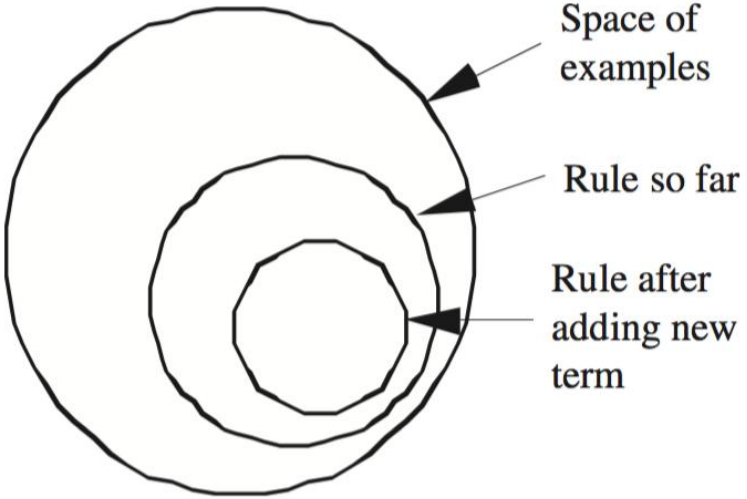
Rules vs. trees

- Z drzewa zawsze można utworzyć zestaw reguł
- zestawy reguł mogą być bardziej przejrzyste, gdy drzewa decyzyjne podlegają replikowanym poddrzewom
- dla problemów wieloklasowych drzewa analizują wszystkie klasy „na raz” a algorytmy pokrywania uwzględnia klasy kolejno

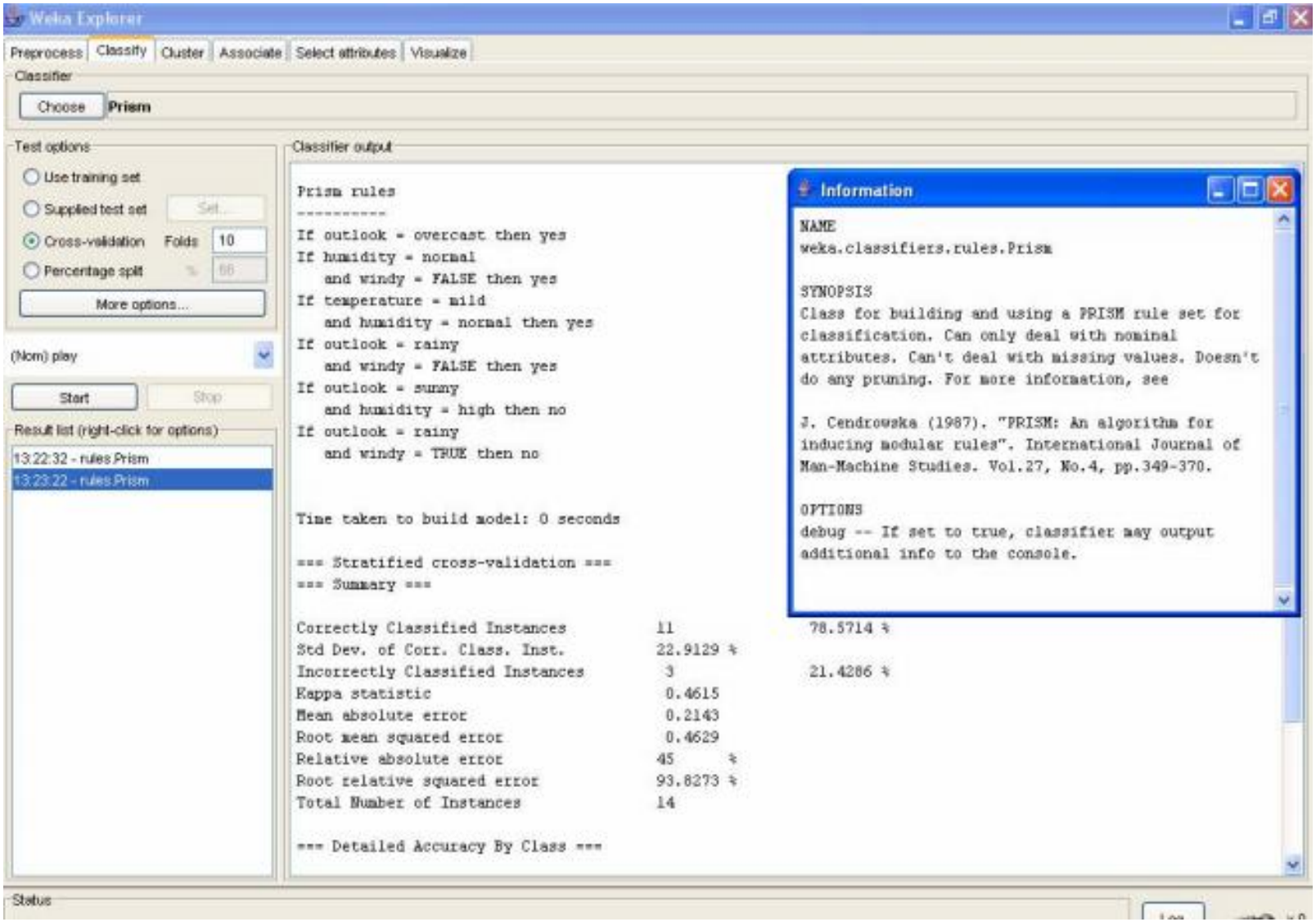


Simple covering algorithm

- Basic idea: generate a rule by adding tests that maximize the rule's accuracy
- Similar to situation in decision trees: problem of selecting an attribute to split on
  - But: decision tree inducer maximizes overall purity
- Each new test reduces rule's coverage:



PRISM - generates a decision list



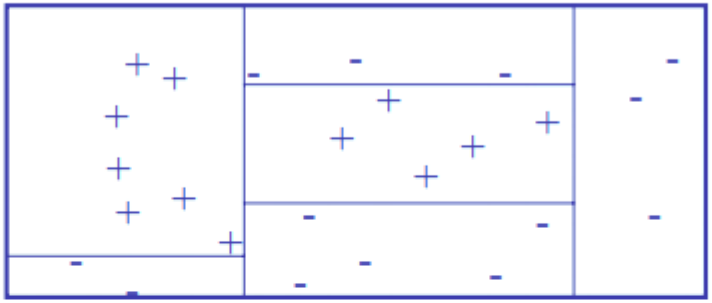
Separate and conquer rule learning

- Rule learning methods like the one PRISM employs (for each class) are called *separate-and-conquer* algorithms:
  - First, identify a useful rule
  - Then, separate out all the instances it covers
  - Finally, “conquer” the remaining instances
- Difference to divide-and-conquer methods:
  - Subset covered by a rule does not need to be explored any further

Decision rules vs. decision trees

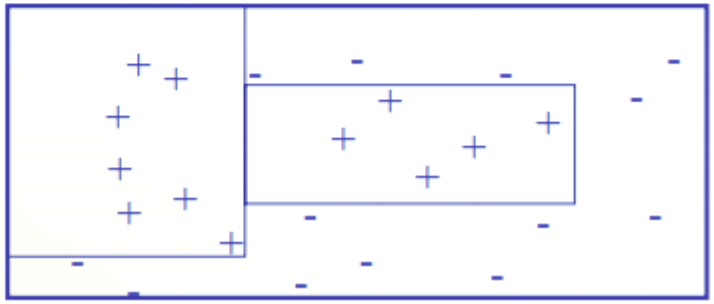
- Trees – splitting the data space (e.g. C4.5)

Decision boundaries of decision trees



- Rules – covering parts of the space (AQ, CN2, LEM)

Decision boundaries of decision rules



Parametry reguł

$a => b$

n_{ab} - liczba obiektów spełniających obie strony reguły

n_a - liczba obiektów spełniających lewą stronę reguły

n_b - liczba obiektów spełniających prawą stronę reguły (czyli w praktyce - liczność klasy decyzyjnej, na którą wskazuje reguła).

/ Dobra reguła to taka, która ma jak najmniej kontrprzykładów, czyli obiektów pasujących do lewej strony, ale niepasujących do prawej.

wsparcie i dokładność

wsparcie (*support*):

$$supp(a => b) = n_{ab} / n$$

wsparcie reguły mówi nam o tym, ile obiektów (treningowych) do niej pasuje.

dokładność (*accuracy*):

$$acc(a => b) = n_{ab} / n_a$$

dokładność reguły mówi nam o jej wiarygodności - jak bardzo możemy liczyć na to, że opisywana przez nią zależność rzeczywiście zachodzi.

	Outlook	Temp.	Humid.	Wind	Sport?
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cold	Normal	Weak	Yes
6	Rain	Cold	Normal	Strong	No
7	Overcast	Cold	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cold	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

Temp=Mild ∧ Humid=Normal ⇒ Sport=Yes

Wsparcie reguły: 0,22

(2 obiekty pasujące, 9 obiektów w klasie Sport=Yes)

Dokładność reguły: 1

Wind=Weak ⇒ Sport=Yes

Wsparcie reguły: 0,66

(6 obiektów pasujących do obu stron, 9 obiektów w klasie Sport=Yes)

Dokładność reguły: 0,75

(6 obiektów pasujących do obu stron, 8 obiektów mających Wind=Weak)

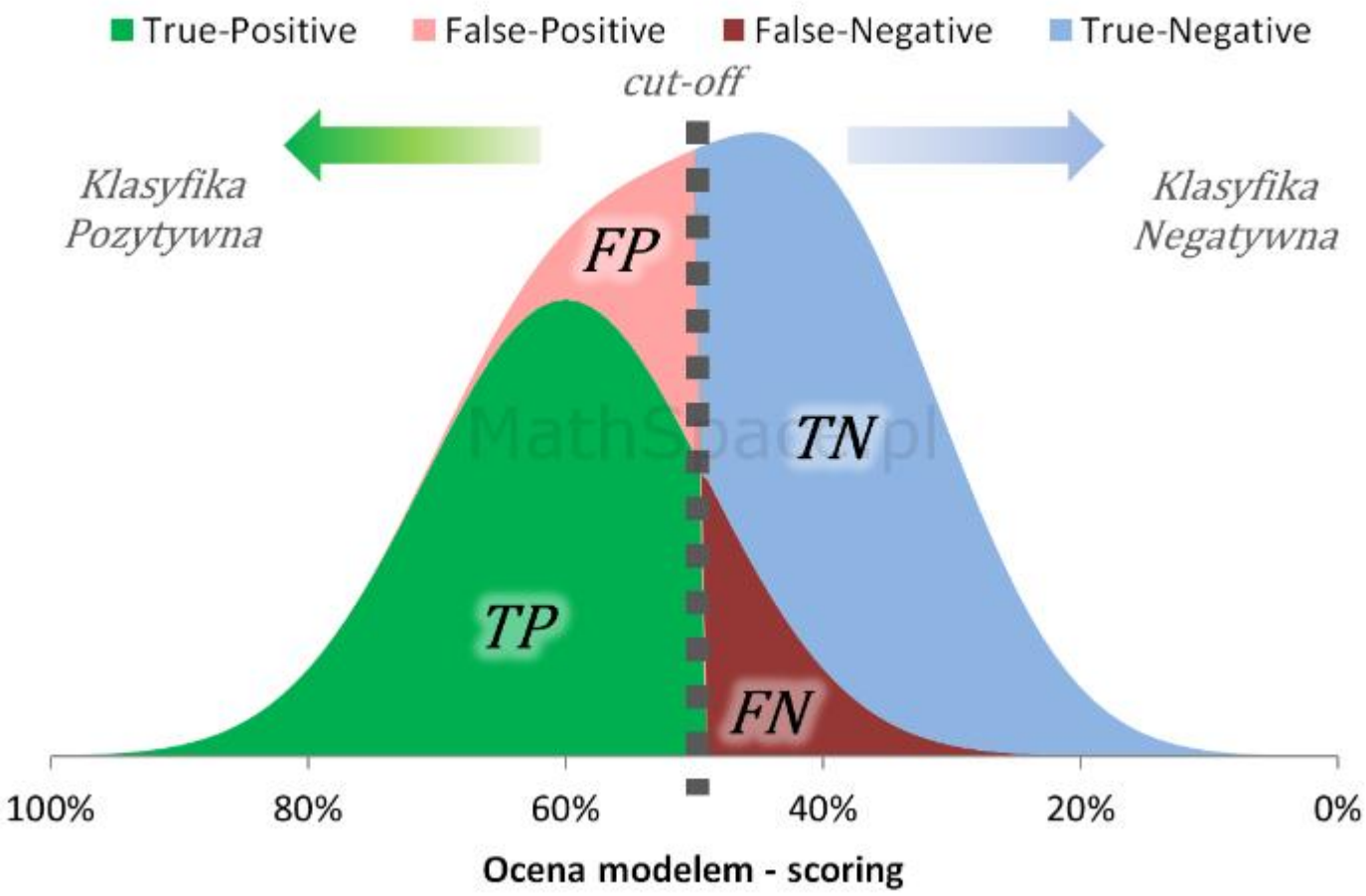
Tablica pomyłek (macierz błędów, macierz klasyfikacji)

Confusion matrix

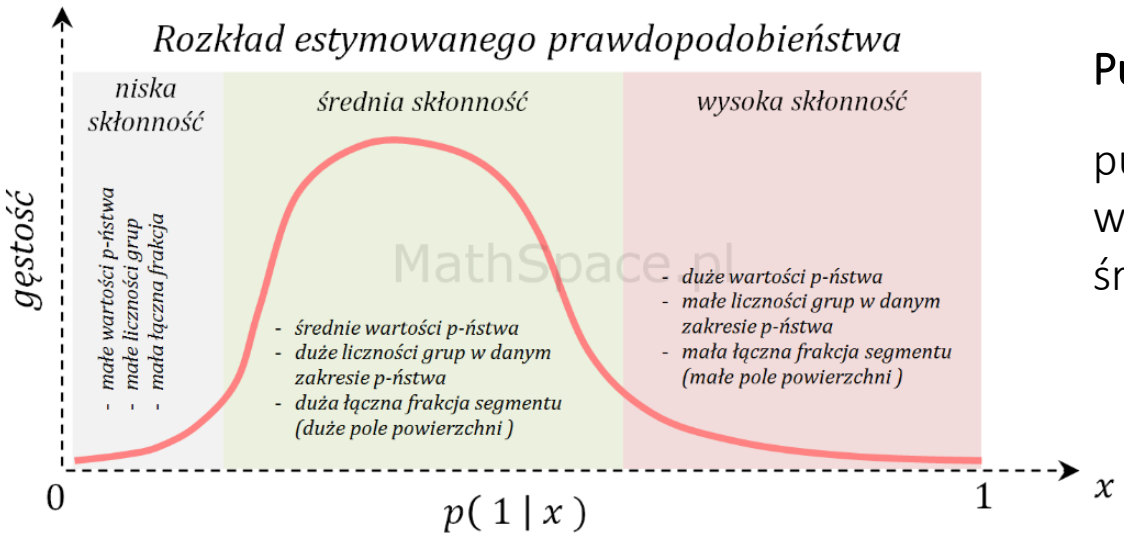
ocena jakości klasyfikacji binarn

		klasa rzeczywista	
		pozytywna	**negatywna**
klasa predykowana	**pozytywna**	prawdziwie pozytywna (TP)	fałszywie pozytywna (FP)
	negatywna	fałszywie negatywna (FN)	prawdziwie negatywna (TN)

		Klasa predykowana – wynik testu			
	Populacja	Klasa pozytywna	Klasa negatywna	Częstość występowania, chorobowość $\frac{\sum \text{stan pozytywny}}{\sum \text{populacja}}$	
Klasa rzeczywista	Stan pozytywny	**Prawdziwie pozytywna**, TP	**Fałszywie negatywna** (błąd drugiego rodzaju, **FN**)	Czułość, TPR $\frac{\sum \text{TP}}{\sum \text{TP}+\sum \text{FN}}$	FNR $\frac{\sum \text{FN}}{\sum \text{TP}+\sum \text{FN}}$
	Stan negatywny	**Fałszywie pozytywna** (błąd pierwszego rodzaju, **FP**)	**Prawdziwie negatywna**, TN	FPR $\frac{\sum \text{FP}}{\sum \text{FP}+\sum \text{TN}}$	Swoistość, SPC, TNR $\frac{\sum \text{TN}}{\sum \text{FP}+\sum \text{TN}}$
	Dokładność, ACC $\frac{\sum \text{TP}+\sum \text{TN}}{\sum \text{populacja}}$	Precyzja, PPV $\frac{\sum \text{TP}}{\sum \text{TP}+\sum \text{FP}}$	FOR $\frac{\sum \text{FN}}{\sum \text{FN}+\sum \text{TN}}$	LR+ $\frac{\text{TPR}}{\text{FPR}}$	DOR $\frac{\text{LR+}}{\text{LR−}}$
		FDR $\frac{\sum \text{FP}}{\sum \text{TP}+\sum \text{FP}}$	NPV $\frac{\sum \text{TN}}{\sum \text{FN}+\sum \text{TN}}$	LR- $\frac{\text{FNR}}{\text{TNR}}$	



ocena jakości klasyfikacji



Punkt odcięcia (cut-off point)

punkt rozgraniczający segment wysokiej skłonności od segmentów średniej i niskiej skłonności

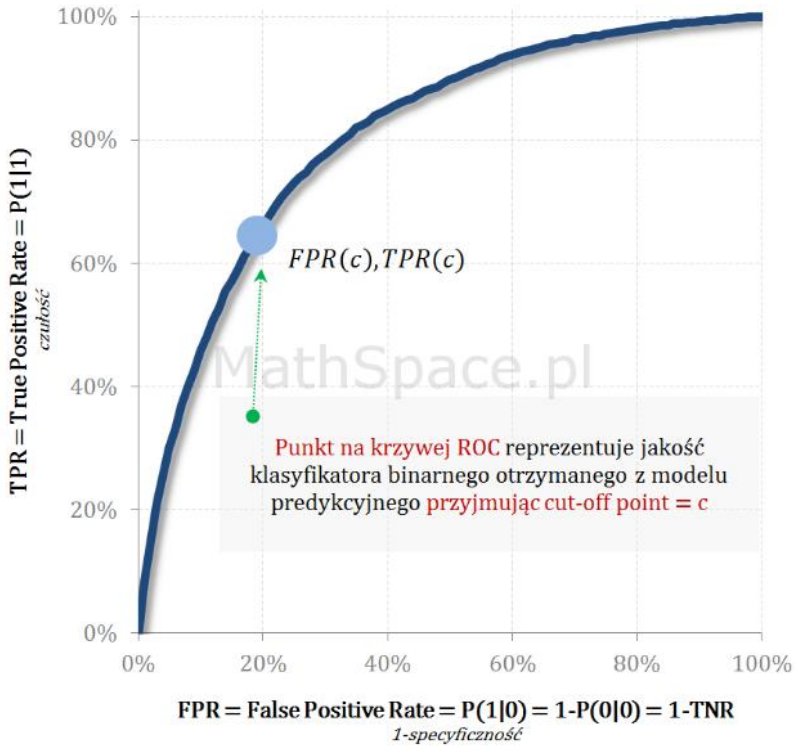
Receiver Operating Characteristic - Krzywa ROC

krzywa ROC jest graficzną reprezentacją efektywności modelu, testujemy klasyfikator dla różnych progów alfa.

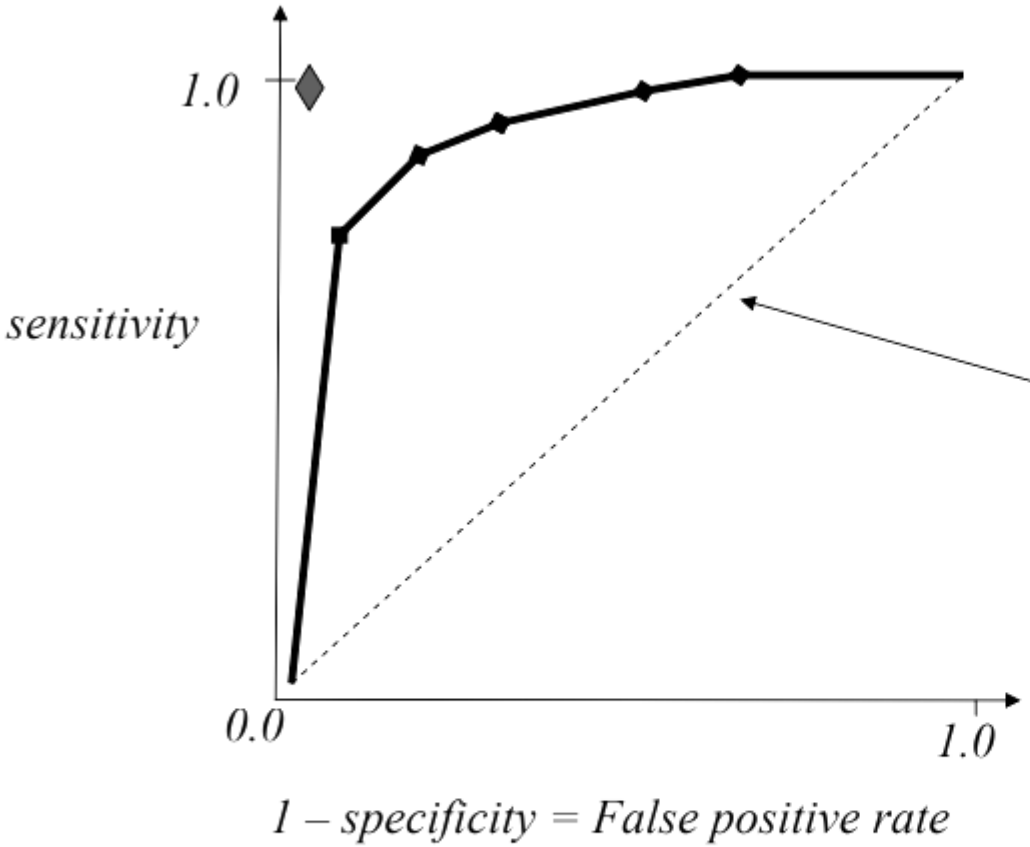
alfa to próg szacowanego prawdopodobieństwa, powyżej którego obserwacja klasyfikowana jest do jednej kategorii (Klasa_pos), a poniżej którego – do drugiej kategorii (Klasa_neg).

ROC pokazuje zależności wskaźników TPR (True Positive Rate) oraz FPR (False Positive Rate).

Im wykres bardziej ”wypukły”, tym lepszy klasyfikator.



ROC



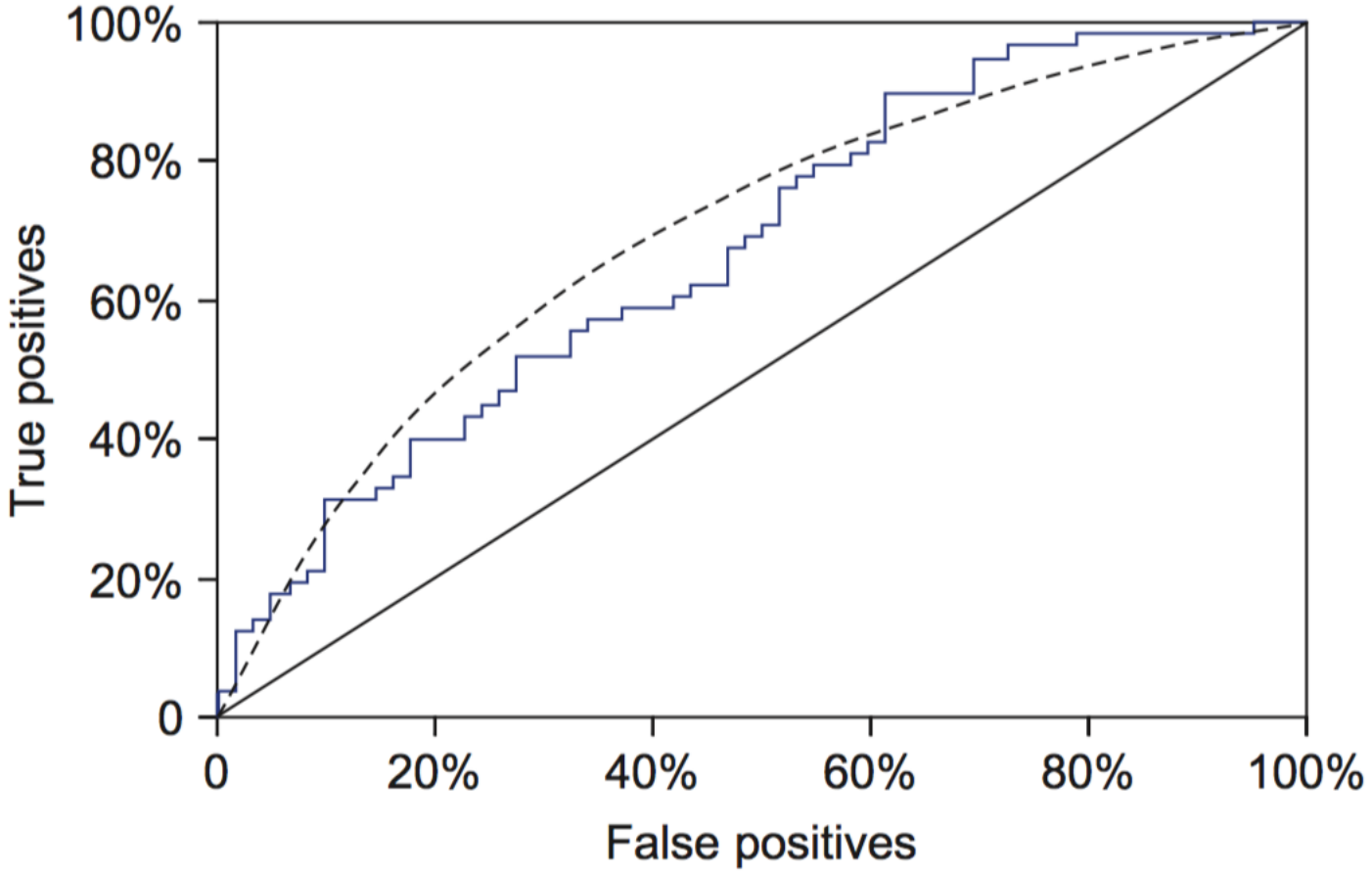
♦ Ideal point .

Diagonal line – random guessing / boundary for useful classifiers

You can compare performance of several classifiers.
Quite often AUC – area under curve – is calculated.

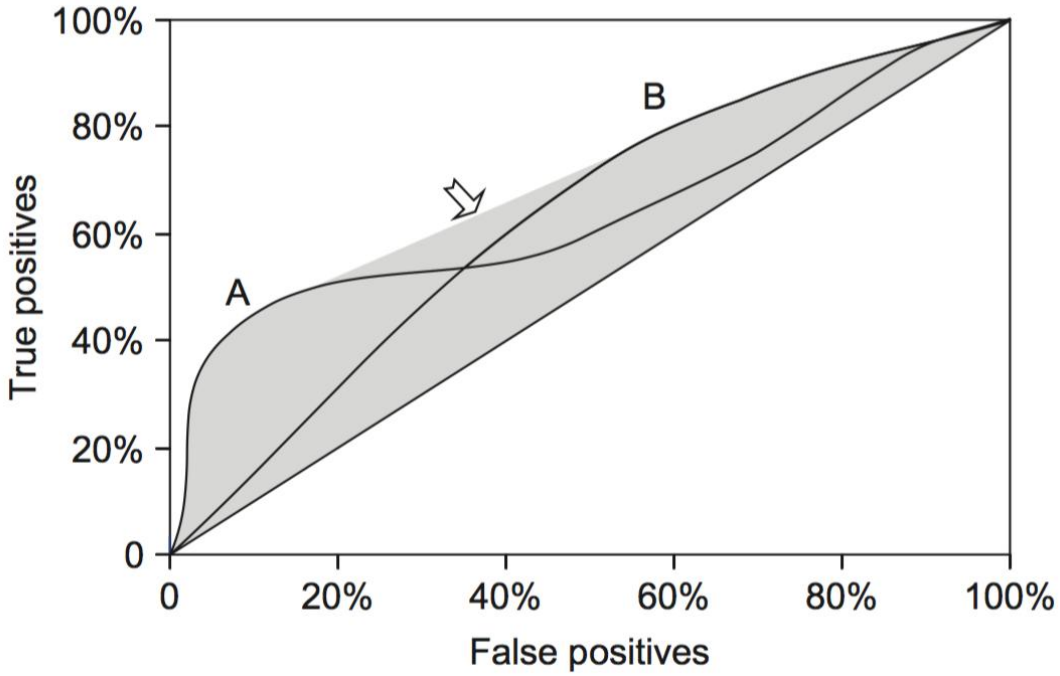
J. Stefanowski

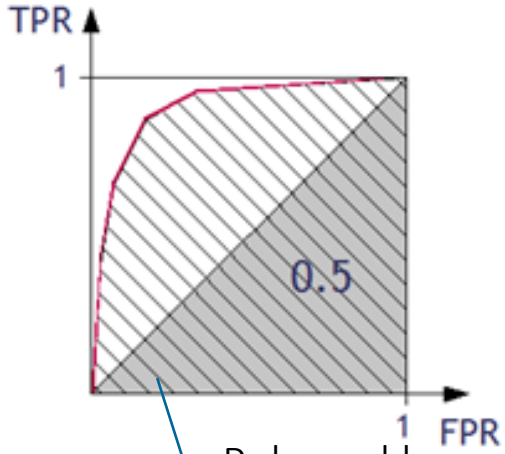
A sample ROC curve



- Jagged curve—one set of test data
- Smoother curve—use cross-validation

ROC curves for two schemes





Pole pod krzywą (AUC) (Area Under ROC Curve)

- For a small, focused sample, use method A
- For a larger one, use method B

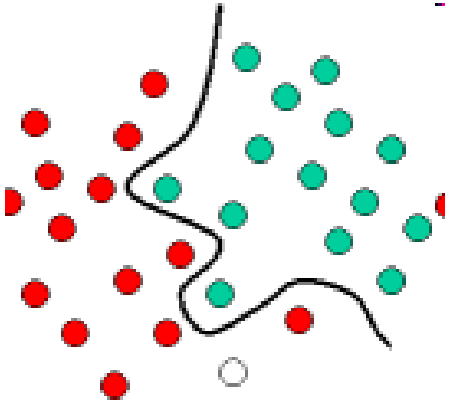
im większe AUC tym lepiej:
AUC = 1 (klasyfikator idealny),
AUC = 0.5 (klasyfikator losowy),
AUC < 0.5 (nieprawidłowy klasyfikator (gorszy niż losowy))

Metoda wektorów nośnych (wspierających)

Support Vector Machines

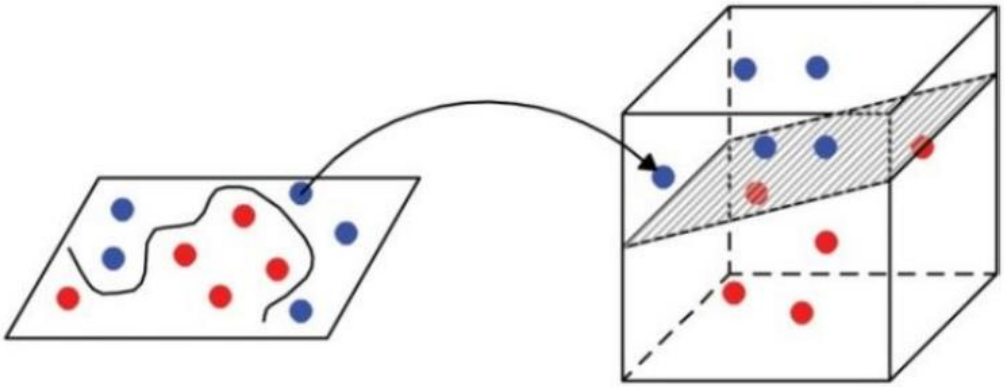
Metoda wektorów nośnych (wspierających)

stosowane gdy do poprawnego klasyfikowania potrzebne są bardziej skomplikowane struktury niż linia prosta



oryginalne obiekty są "mapowane" (transformowane) za pomocą funkcji jądrowych (kernels) na przestrzeń ilustrowaną po prawej.

w nowej przestrzeni dwie klasy są liniowo separowalne, co pozwala uniknąć skomplikowanej postaci granicy klas.



Problem oryginalny nieseparowalny liniowo w 2D

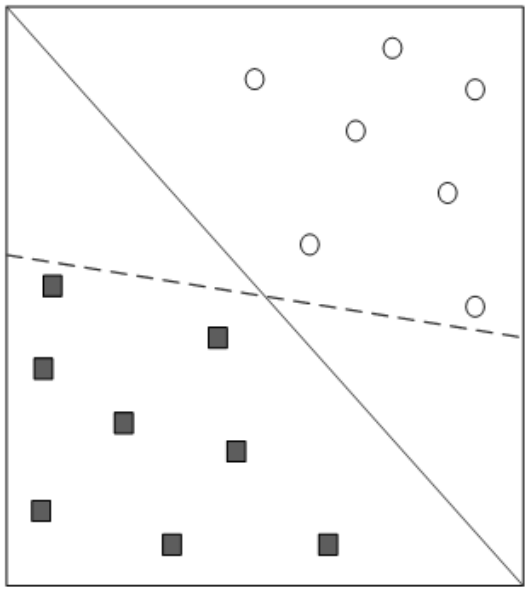
Odwzorowanie poblemu oryginalnego w przestrzeni 3D separowalne liniowo

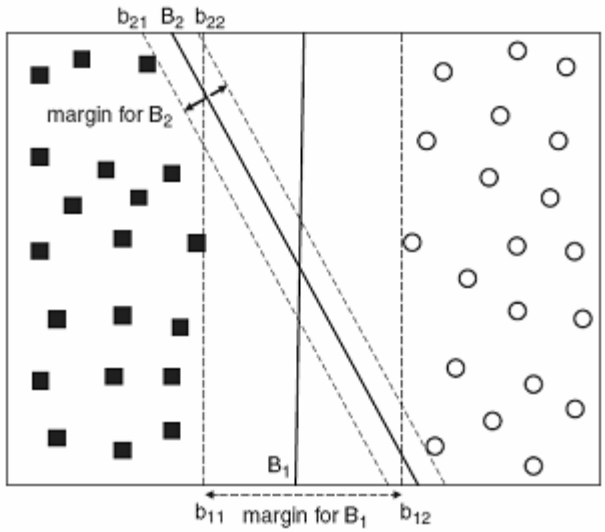
Support Vector Machines

Którą z hiperpłaszczyzn należy wybrać? B1 or B2?

Hiperpłaszczyzny b_{i1} i b_{i2} są otrzymane przez równoległe przesuwanie hiperpłaszczyzny granicznej aż do pierwszych punktów z obu klas.

Odległość między nimi – margines klasyfikatora liniowego

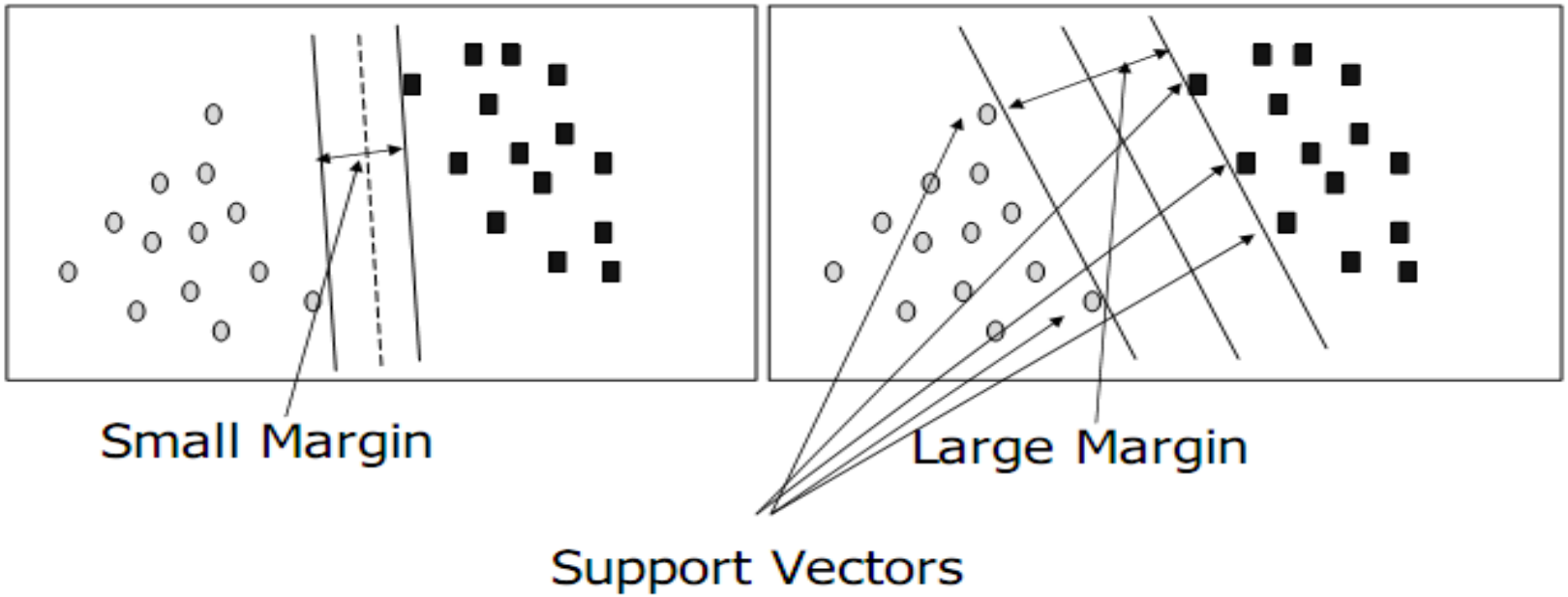




Węższe czy szersze marginesy?

Szerszy margines → lepsze własności generalizacji, mniejsza podatność na ew. przeuczenie (overfitting)

Wąski margines – mała zmiana granicy, radykalne zmiany klasyfikacji

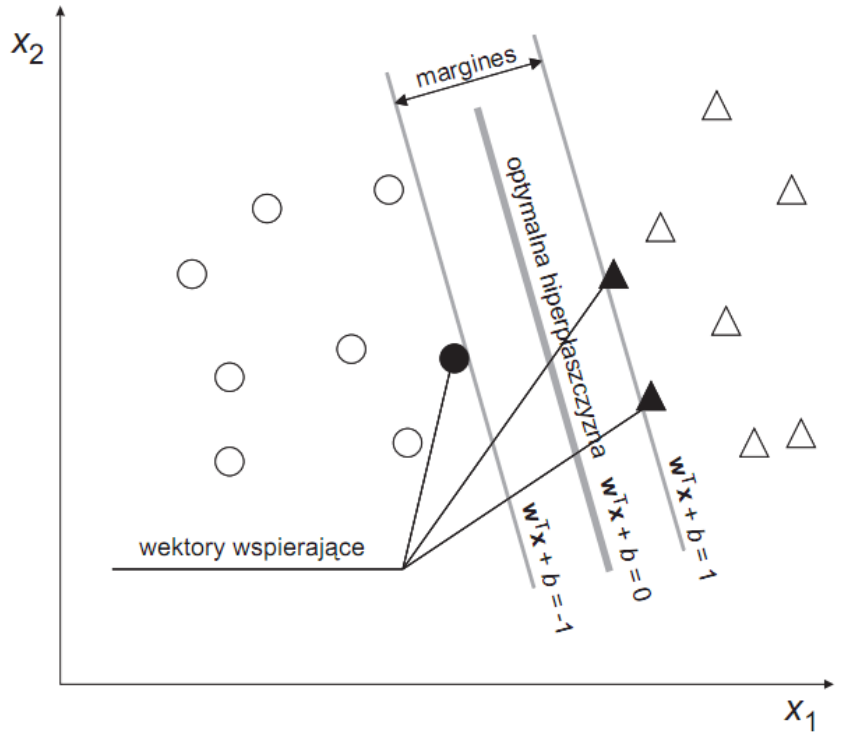


Cel

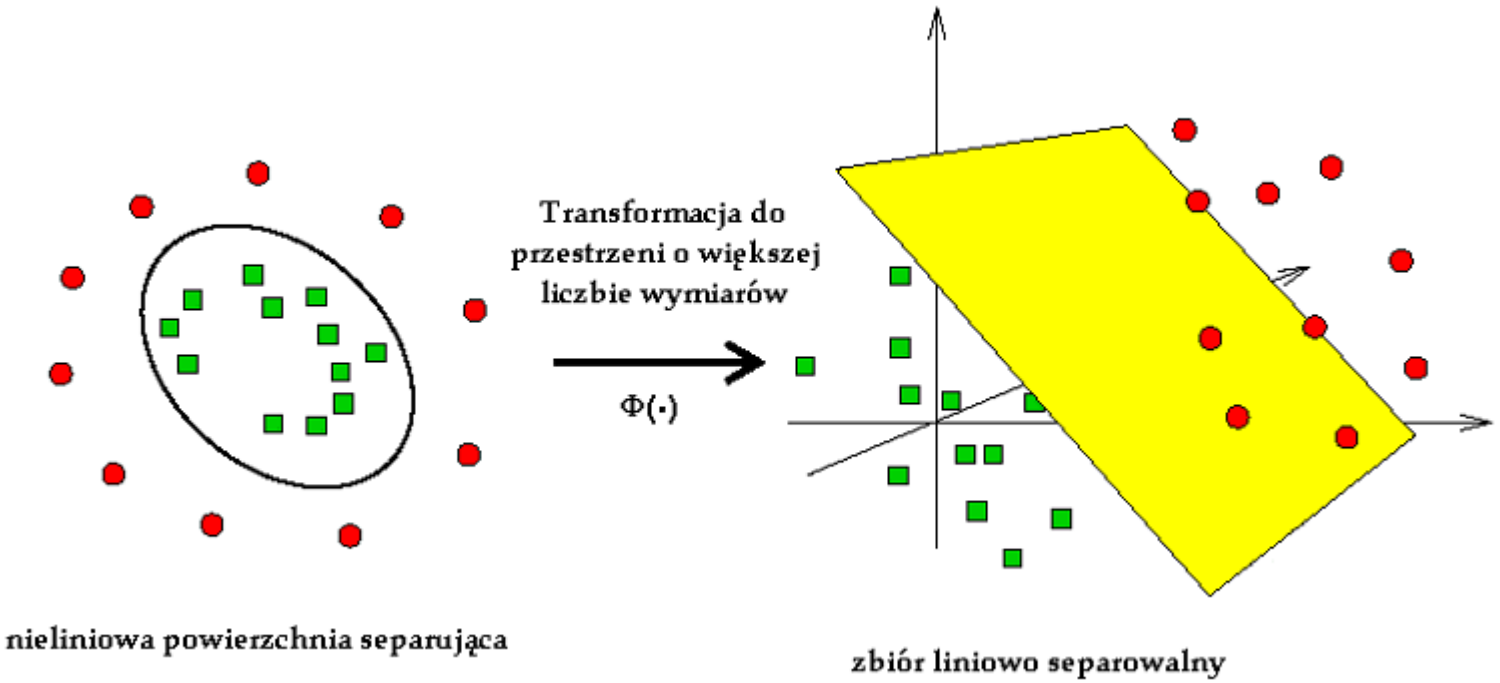
Znajdź hiperpłaszyczną, która maksymalizuje margines
=> B1 jest lepsze niż B2

Maksymalizując odległość pomiędzy B1 i B2 (przy założeniu, że pomiędzy tymi hiperpłaszczyznami nie ma punktów) doprowadzamy do sytuacji kiedy na wspomnianych hiperpłaszczyznach znajdą się punkty należące do zbioru treningowego.

Punkty te nazywane są wektorami nośnymi, ponieważ tylko one uczestniczą w definicji hiperpłaszczyzn separujących.



Funkcje jądrowe



Iloczyn skalarny w przestrzeni o większym wymiarze jest równoważny funkcji jądra w przestrzeni oryginalnej.

Tak więc nie musimy znać jawnej postaci przekształcenia Φ, wystarczy, że znamy funkcję jądra

Funkcje jądrowe

Istnieje wiele takich funkcji nę

- liniowe:

$$K(a_i, a_j) = a_i^T a_j$$

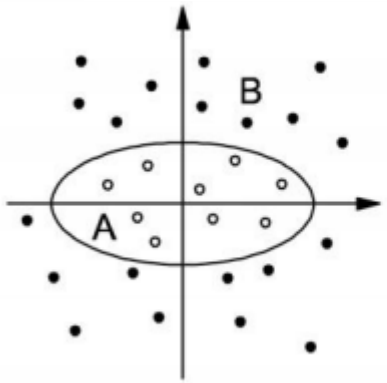
- wielomianowe rzędu d:

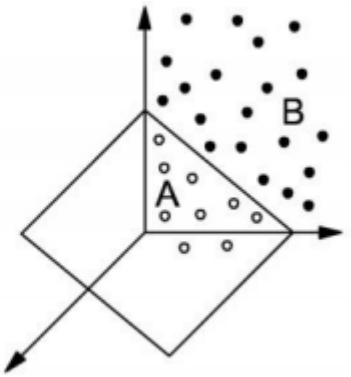
$$K(a_i, a_j) = (a_i^T a_j + 1)^d$$

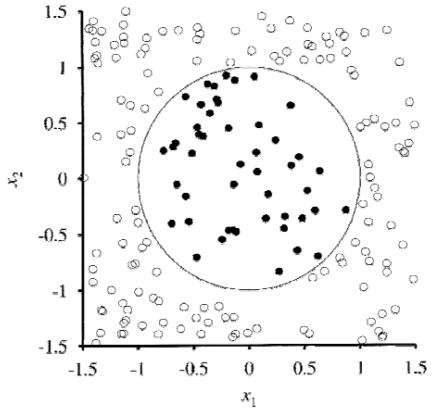
- radialne funkcje bazowe (RBF):

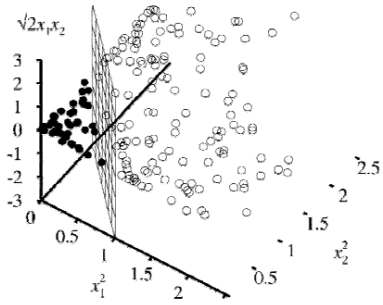
$$K(a_i, a_j) = e^{-\|a_i - a_j\|^2 / 2\sigma^2}$$

Problem klasyfikacji separowalny za pomocą elipsy w przestrzeni oryginalnej

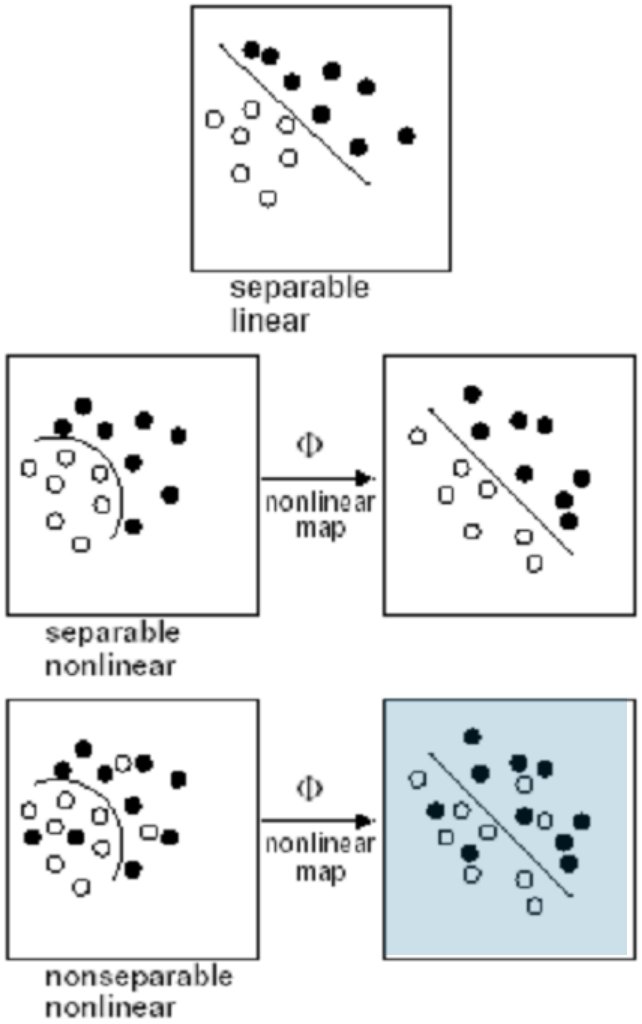








klasyfikacja C-SVM



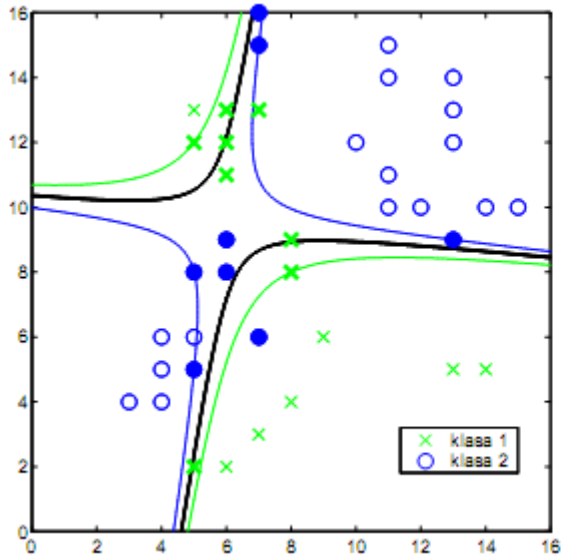
W praktyce często zdarza się, że niemożliwe jest idealne odseparowanie obiektów należących do poszczególnych klas.

Dopuszczamy aby pomiędzy hiperpłaszczyznami H1 i H2 pojawiły się punkty. Jednak, każdy taki punkt jest „karany”.

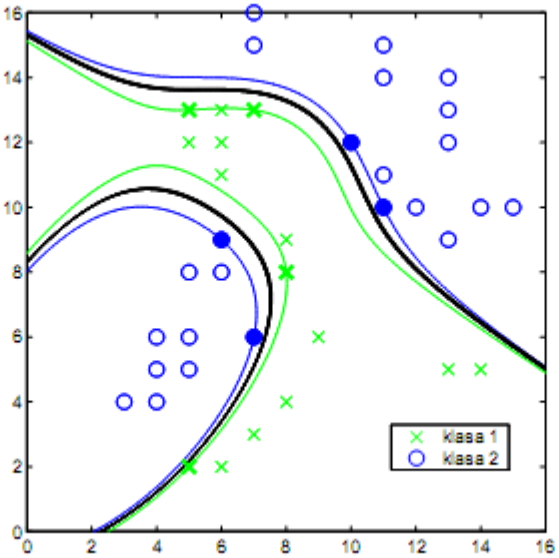
Wprowadza się określenie współczynnika k[illegible] mamy przypade[illegible] separacji klas)

$$\min_{w,b,\xi_i} \frac{1}{2} w^T w + C \sum_{i=1}^{N} \xi_i$$

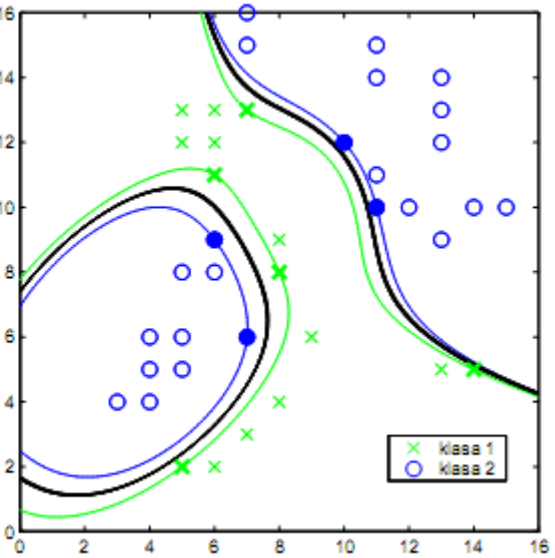
FUNKCJE JĄDRA



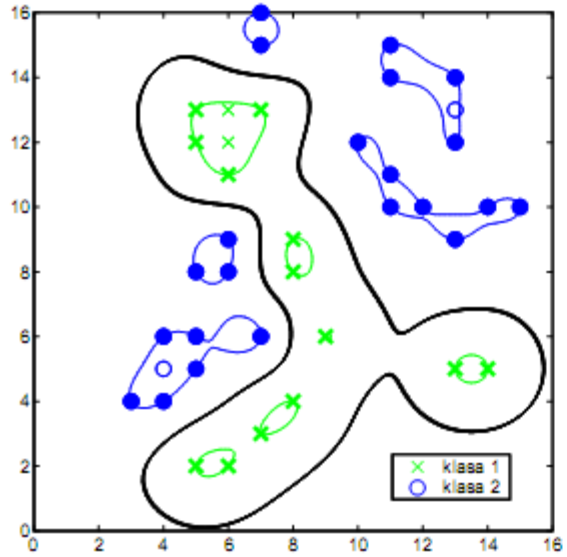
wielomian 2-stopnia



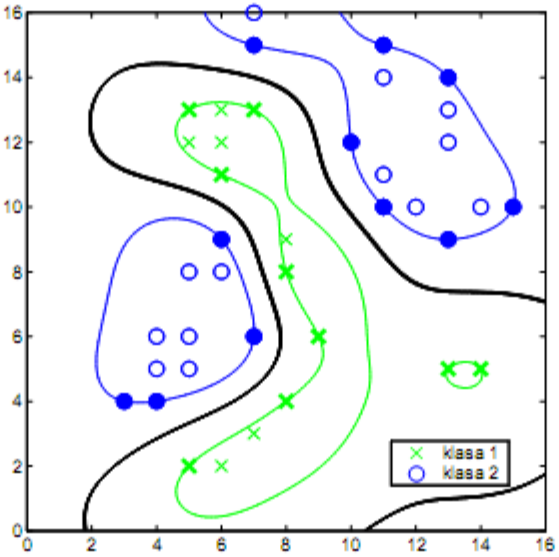
wielomian 3-stopnia



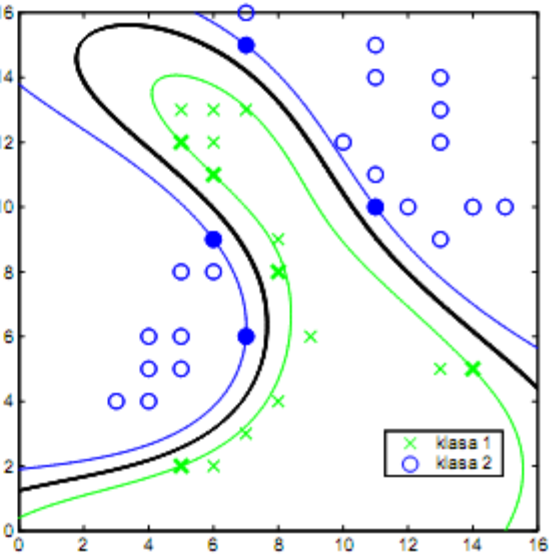
wielomian 4-stopnia



funkcja radialna σ = 1.0

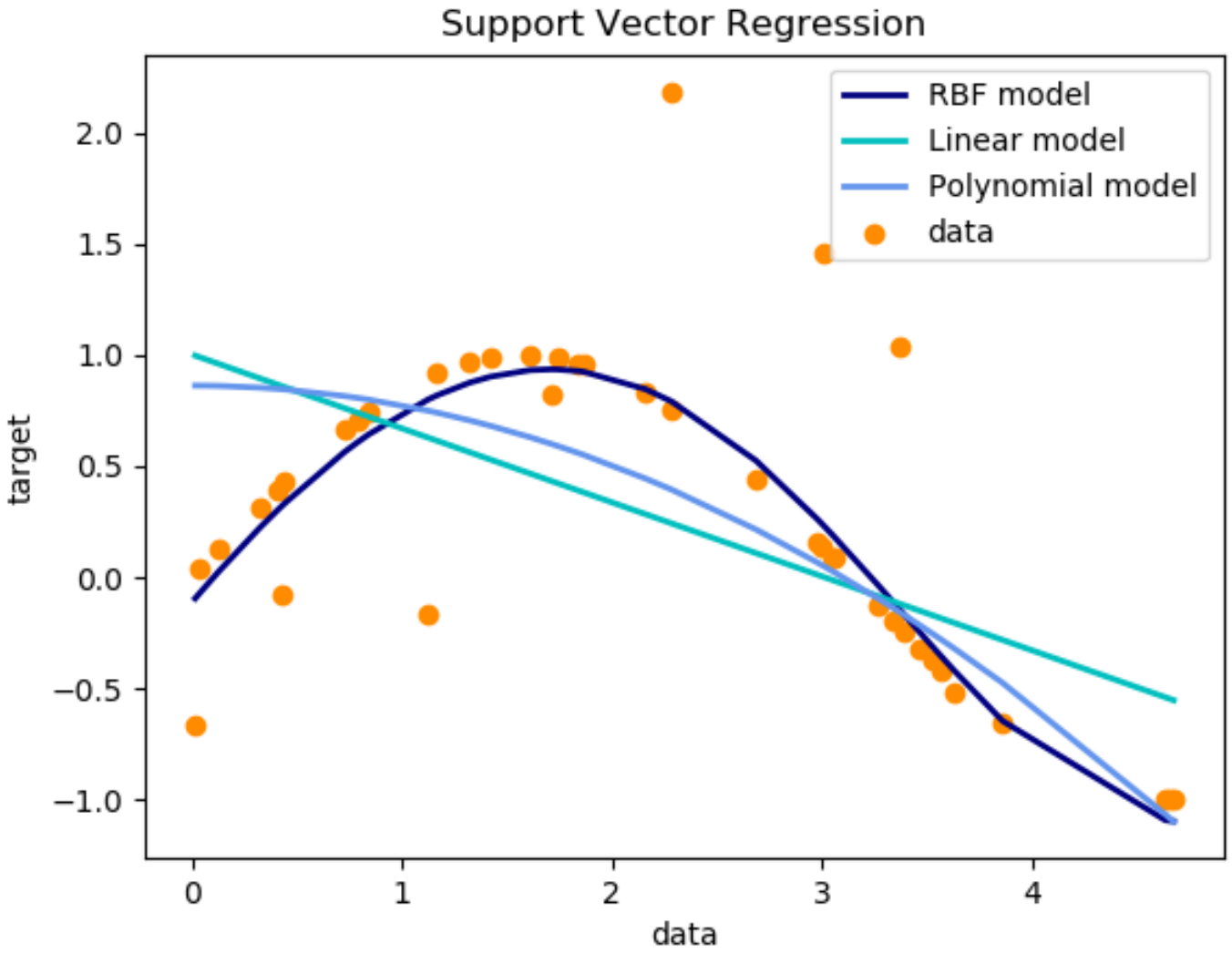


funkcja radialna σ = 2.0



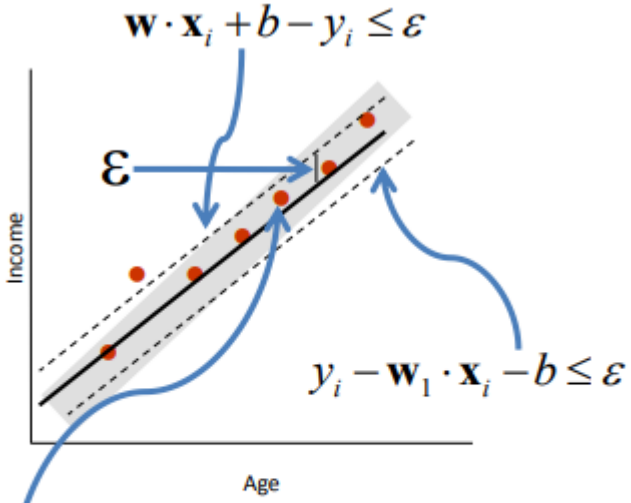
funkcja radialna σ = 5.0

Support Vector Regression (SVR)



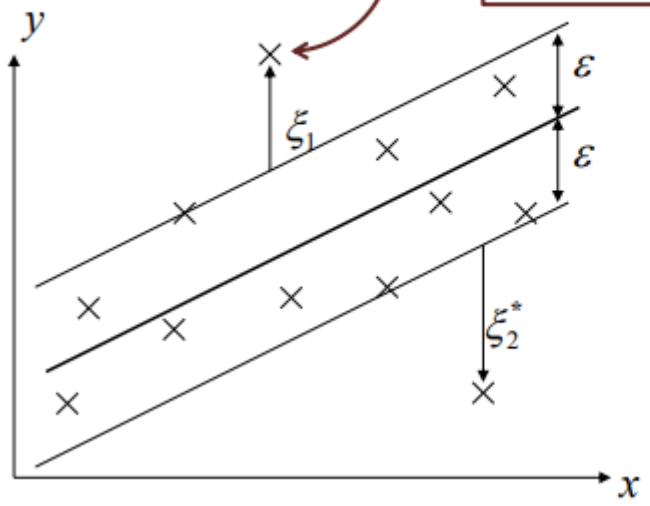
Support Vector Regression

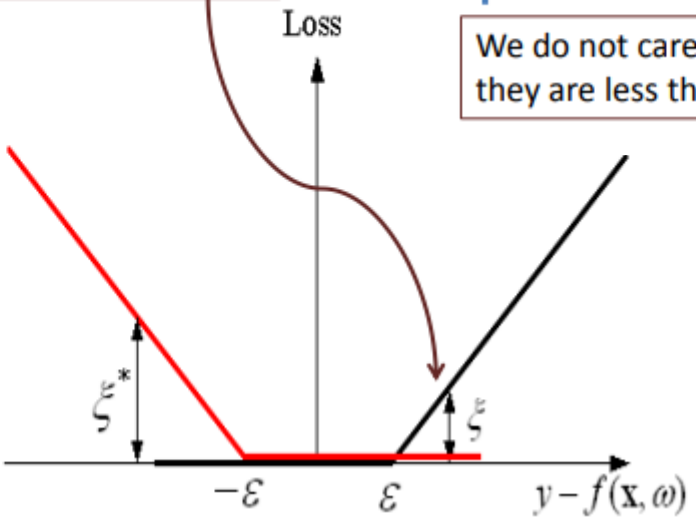
Find a function, f(x), with at most ε-deviation from the target y



Only the point outside the ε-region contribute to the final cost

We do not care about errors as long as they are less than ε





ε-deviation

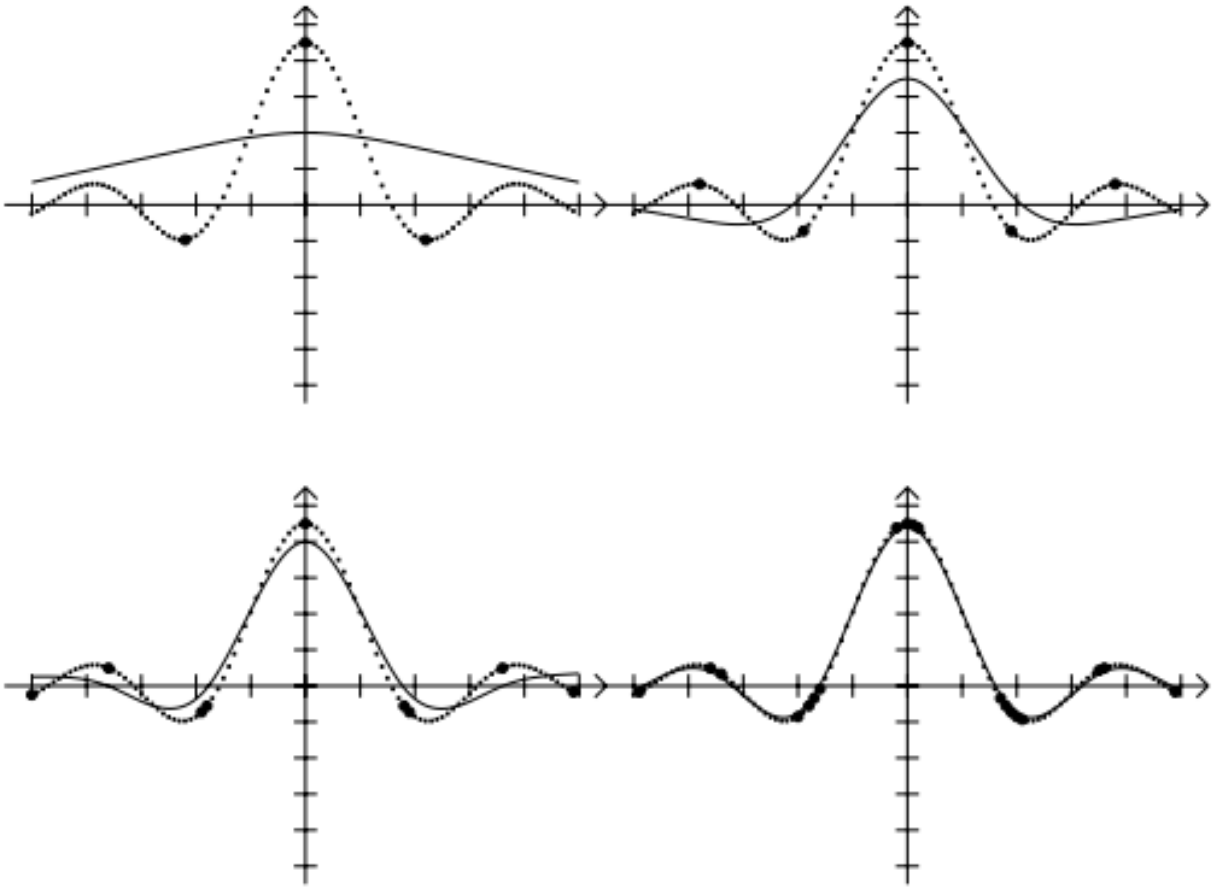
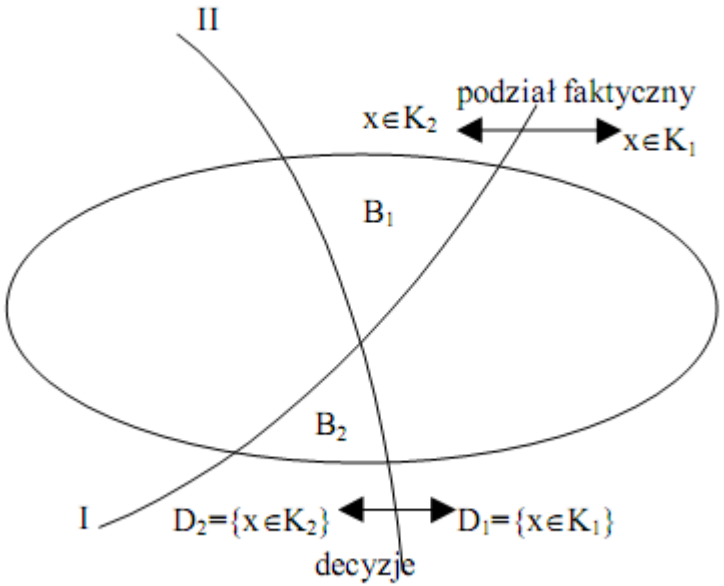


Figure 5 Upper left: regression (solid line), datapoints (small dots) and SVs (big dots) for an approximation with $\varepsilon = 0.5$, upper right $\varepsilon = 0.2$, lower left $\varepsilon = 0.1$, lower right $\varepsilon = 0.02$. Note the increase in the number of SVs.

Analiza dyskryminacyjna

Struktura rzeczywista a klasyfikacja

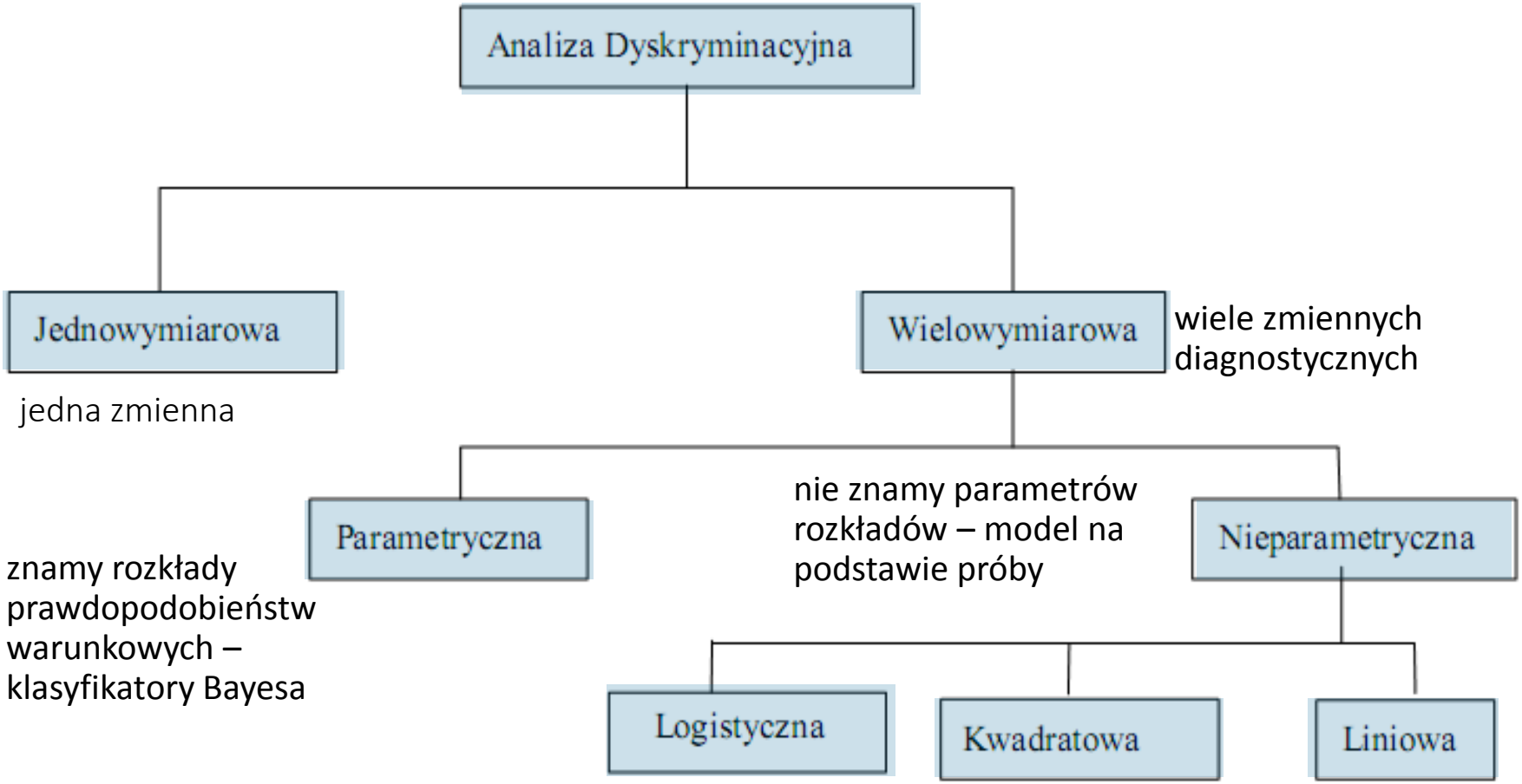


B_1 i B_2 to podzbiory błędnych decyzji

miara obszaru błędnych decyzji charakteryzuje procedurę klasyfikacyjną

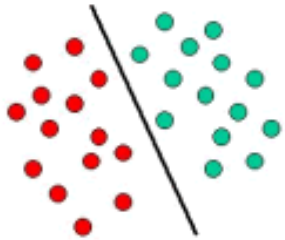
reguły klasyfikacyjne różnią się metodą mierzenia błędu

Klasyfikacja metod klasyfikacji



Liniowa funkcja separująca (graniczna)

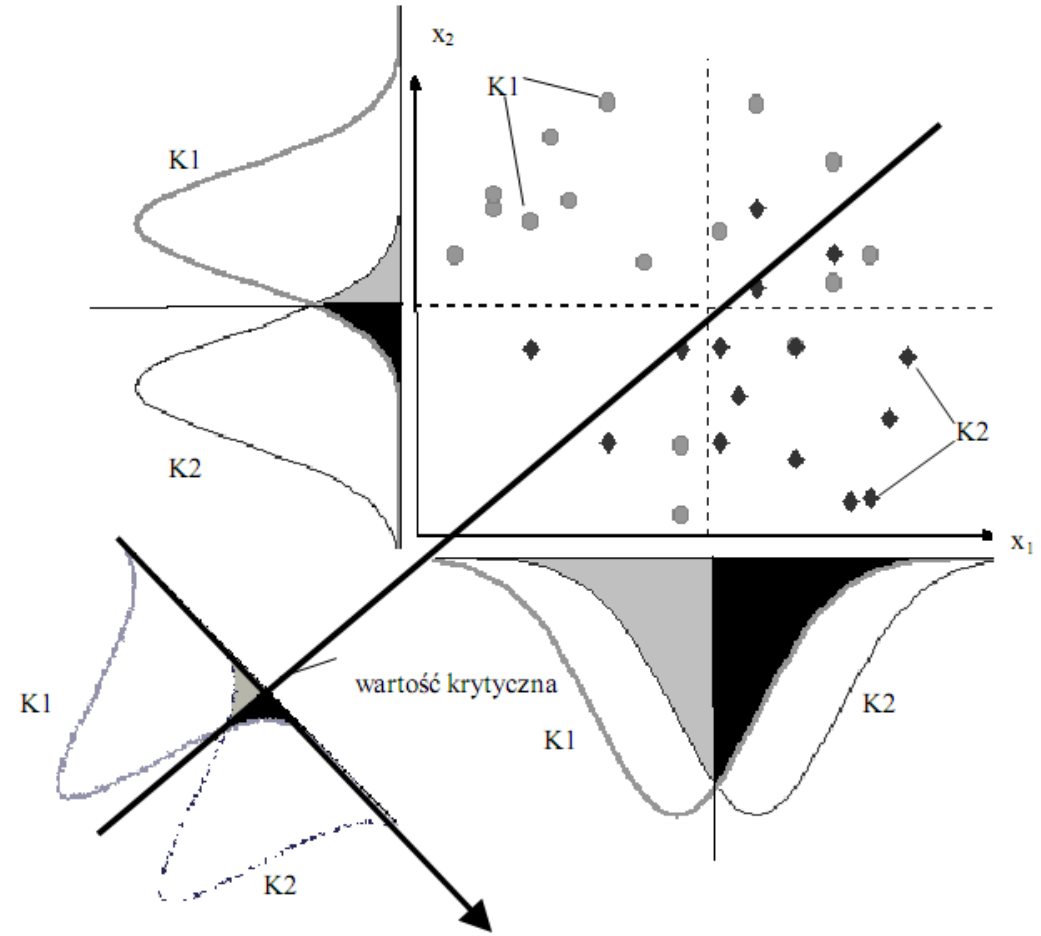
Szukamy klasyfikatora pozwalającego na podział całej przestrzeni na obszary odpowiadające klasom (dwóm lub więcej) oraz pozwalającego jak najlepiej klasyfikować nowe obiekty x do klas



Podejście opiera się na znalezieniu tzw. granicy decyzyjnej między klasami

$f(x)=w^T \cdot x$

LDA



Legenda:
●-obiekty pierwszej klasy;
♦- obiekty drugiej klasy;
- błąd I-ego rodzaju;
- błąd II-ego rodzaju.

Różne podejścia do budowy klasyfikatorów liniowych

Podejścia generatywne (probabilistyczne)

- Analiza dyskryminacyjna (związana z rozkładem normalnym)
- Wersja klasyfikacji Bayesowskiej (dwumianowy rozkład)

Podejścia wykorzystujące własności zbioru uczącego

- Perceptron liniowy Rosenblata (iteracyjne poprawki wag)
- Metoda wektorów nośnych (max. marginesu klasyfikatora)
- Regresja logistyczna (EM estymacja)

Funkcje klasyfikacyjne

Funkcje klasyfikacyjne mogą być wykorzystane do rozstrzygania, do której grupy najprawdopodobniej należą poszczególne przypadki.

Jest tyle funkcji klasyfikacyjnych ile grup. Każda funkcja pozwala nam obliczyć *wartości klasyfikacyjne* dla każdego przypadku w każdej grupie, przy pomocy wzoru:

$$S_i = c_i + w_{i1}*x_1 + w_{i2}*x_2 + \ldots + w_{im}*x_m$$

gdzie,
indeks i określa daną grupę; indeksy $1, 2, \ldots, m$ określają m zmiennych;
c_i jest stałą dla i-tej grupy,
w_{ij} jest wagą dla j-tej zmiennej przy obliczaniu wartości klasyfikacyjnej dla i-tej grupy;

Prawdopodobieństwa klasyfikacyjne *a priori*

Czasami wiemy, że w jednej z grup jest więcej obserwacji niż w jakiejś innej; zatem prawdopodobieństwo *a priori*, że przypadek należy do tej grupy, jest większe.

Należy zadać sobie pytanie, czy nierówna liczba przypadków w różnych grupach w próbie jest odzwierciedleniem rzeczywistego rozkładu w populacji, czy jest to tylko efekt losowania.

Na przykład, jeśli wiemy, że 60% absolwentów szkoły średniej zwykle wstępuje na studia (20% idzie do szkoły pomaturalnej, a pozostałe 20% do pracy), to nasze przewidywanie: *a priori* przy takich samych pozostałych warunkach, jest bardziej prawdopodobne, że uczeń pójdzie na studia, niż że wybierze którąś z pozostałych możliwości.

Analiza dyskryminacyjna umożliwia określenie różnych prawdopodobieństw *a priori*, które zostaną następnie wykorzystanie do skorygowania klasyfikacji przypadków (i obliczenia prawdopodobieństw *a posteriori*).

Macierz klasyfikacji

W celu rozstrzygnięcia, na ile dobrze bieżące funkcje klasyfikacyjne pozwalają przewidzieć przynależność przypadków do grupy oglądamy *macierz klasyfikacji*.

Macierz klasyfikacji pokazuje liczbę przypadków, które zostały poprawnie sklasyfikowane (na przekątnej macierzy) oraz tych, które zostały błędnie zaklasyfik

Macierz klasyfikacji (Gminy1.sta)
Wiersze: obserwowana klasyfik.
Kolumny: Przewidywana klasyfikacja

Grupa	**Procent Poprawne**	G_1:1 p=,33242	G_2:2 p=,33379	G_3:3 p=,33379
G_1:1	85,53719	207	3	32
G_2:2	75,72016	6	184	53
G_3:3	64,19753	29	58	156
Razem	75,13736	242	245	241

Analiza dyskryminacyjna

jest stosowana do rozstrzygania, które zmienne pozwalają w najlepszy sposób dzielić dany zbiór przypadków na klasy.

główna idea analizy funkcji dyskryminacyjnej to rozstrzyganie, czy grupy różnią się ze względu na średnią pewnej zmiennej, a następnie wykorzystanie tej zmiennej do przewidywania przynależności do grupy (np. nowych przypadków).

np. w badaniach medycznych można rejestrować różne zmienne związane ze stanem zdrowia pacjentów, aby sprawdzić, które zmienne najlepiej prorokują, czy pacjent ma szansę na zupełne wyleczenie (grupa 1), częściowe wyleczenie (grupa 2), czy nie ma szans (grupa 3) na wyleczenie.

Podejście obliczeniowe

Z rachunkowego punktu widzenia, analiza funkcji dyskryminacyjnej jest bardzo podobna do analizy wariancji (ANOVA).

np. mierzymy wzrost w losowej próbie 50 mężczyzn i 50 kobiet. Kobiety nie są, przeciętnie, tak wysokie jak mężczyźni, a różnica ta znajdzie odbicie w różnicy średnich (dla zmiennej *Wzrost*). Dlatego zmienna wzrost pozwala nam zróżnicować mężczyzn i kobiety z *większym niż przypadkowe* prawdopodobieństwem:

jeśli osoba jest wysoka, to prawdopodobnie jest mężczyzną, jeśli osoba jest niska, to prawdopodobnie jest kobietą.

ANOVA

zagadnienie funkcji dyskryminacyjnej może być przeformułowane na problem jednoczynnikowej analizy wariancji (*ANOVA*).

można zapytać, czy dwie (lub więcej) grupy różnią się istotnie od siebie ze względu na średnią pewnej zmiennej.

jeśli średnie pewnej zmiennej są istotnie różne w różnych grupach, to możemy powiedzieć, że ta zmienna dyskryminuje te grupy.

aby rozstrzygnąć, czy są jakieś istotne różnice (odnośnie wszystkich zmiennych) między grupami, możemy porównać macierze całkowitych wariancji i kowariancji przy pomocy wielowymiarowych testów F.

Założenia analizy dyskryminacyjnej

Zmienne wyrażone na skalach liczbowych

» Specjalne podejścia dla zmiennych jakościowych (binaryzacja, model lokacyjny,…)

Rozkład normalny. Zakłada się, że zmienne reprezentują próbę z wielowymiarowego rozkładu normalnego.

przy pomocy analizy dyskryminacyjnej bardzo łatwo można tworzyć histogramy rozkładów liczebności.

naruszanie założenia o normalności zazwyczaj nie jest "zgubne" w tym sensie, że wypadkowe testy istotności itd. pozostają odporne. W module ANOVA/MANOVA znajdują się specjalne testy na normalność rozkładu.

Założenia analizy dyskryminacyjnej

Korelacje między średnimi i wariancjami.
Podstawowe "rzeczywiste" zagrożenie dla trafności testów istotności pojawia się wówczas, gdy średnie zmiennych w grupach są skorelowane z wariancjami (lub odchyleniami standardowymi).

Ogólne testy istotności są oparte na zgrupowanych wariancjach, to znaczy na przeciętnej wariancji z wszystkich grup, odbijając się na istotności statystycznej.

W praktyce, model taki może się pojawić wtedy, gdy jedna z badanych grup zawiera kilka przypadków odstających, które mają duży wpływ na średnie a także zwiększają zmienność.

Aby ustrzec się przed tym problemem, skontrolujmy

Założenia analizy dyskryminacyjnej

Problem złego uwarunkowania macierzy wymaga, by zmienne wykorzystywane do dyskryminacji grup nie były w pełni redundantne.

Częścią obliczeń analizy dyskryminacyjnej jest odwrócenie macierzy wariancji/kowariancji zmiennych w modelu.

Jeśli któraś ze zmiennych jest redundantna wobec innych zmiennych, to o macierzy mówi się, że jest źle uwarunkowana i nie może być odwrócona.

Na przykład, jeśli zmienna jest sumą trzech innych zmiennych, które także znajdują się w modelu, to macierz jest źle uwarunkowana.

Założenia analizy dyskryminacyjnej

Wartości tolerancji. Aby ustrzec się złego uwarunkowania macierzy *można* sprawdzać dla każdej zmiennej tak zwaną wartość tolerancji.

Wartość tolerancji jest obliczana jako $1 - R^2$ danej zmiennej przy włączeniu do bieżącego modelu wszystkich innych zmiennych. Jest to *część wariancji wyjaśniana przez zmienną*.

gdy zmienna jest prawie zupełnie redundantna (a zatem może pojawić się problem złego uwarunkowania macierzy), wartość tolerancji dla tej zmiennej zbliży się do 0.

Domyślna wartość w analizie dyskryminacyjnej dla minimalnej akceptowalnej tolerancji wynosi 0.01. Gdy tolerancja dla dowolnej zmiennej wypadnie poniżej tej wartości, to znaczy, że zmienna będzie redundantna w więcej niż 99%

Liniowa analiza dyskryminacyjna

Problem wprowadzony przez R.A. Fishera w 1936 dla wielowymiarowej przestrzeni atrybutów (zmiennych liczbowych) – dyskryminacja 2 klas

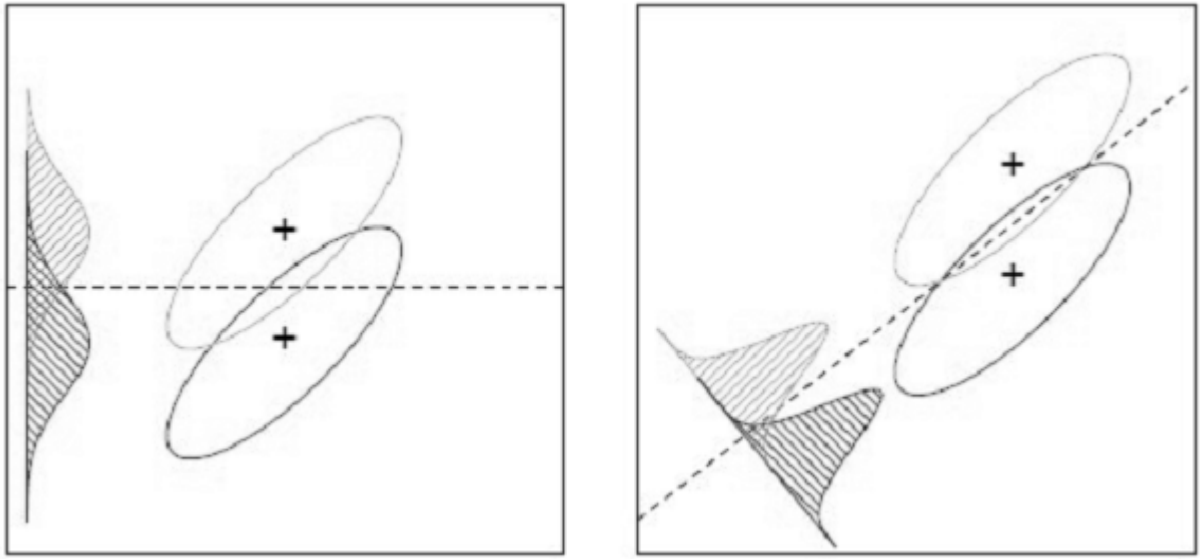
Fisher oryginalnie zaproponował poszukiwanie kierunku projekcji, na którym można dobrze rozdzielić zrzutowane obie klasy

- Średnie w klasach są dostatecznie oddalone od siebie
- Obszary rozrzutu (rozproszenia, zmienności) obu klas nie nakładają się zbyt mocno.

LDF – Linear Discriminant Function

FLD – Fisher Linear Discriminant

Liniowa analiza dyskryminacyjna



Rzutowanie na linię łączącą środki gęstości dałoby największy rozrzut środków gęstości, ale wartości nachodzą na siebie ze względu na kowariancję.

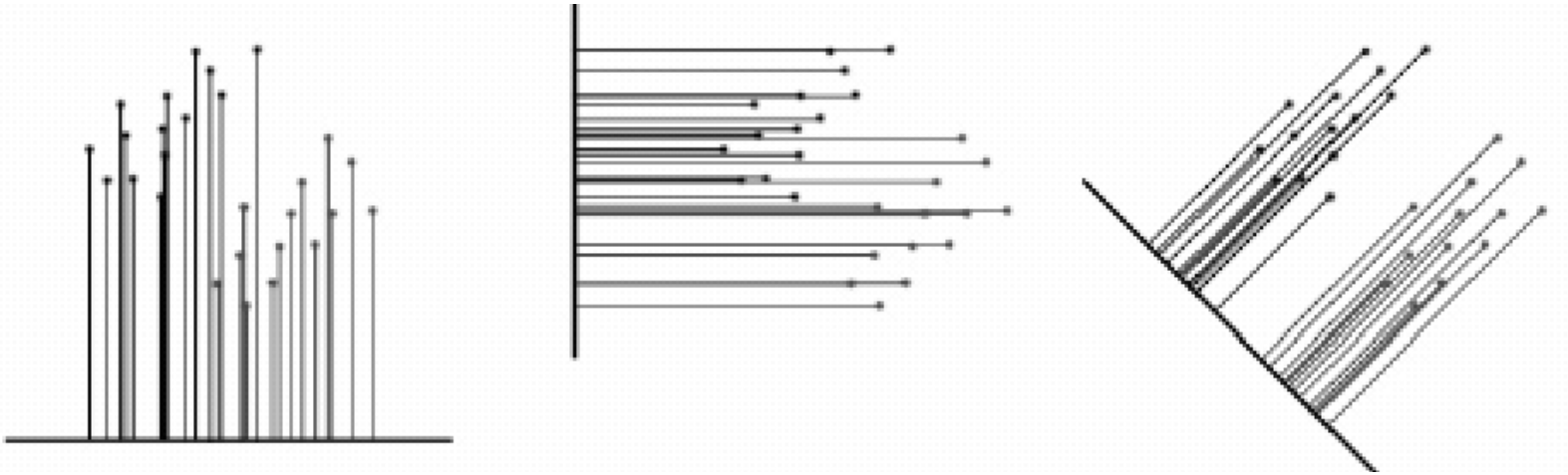
dyskryminujący kierunek projekcji pozwala zmniejszyć to nakładanie dla zmiennych o rozkładzie normalnym

na podstawie zbioru uczącego szukamy takiego kierunku, dla którego seperacja danych na klasy jest najwieksza, a „zachodzenie” najmniejsze

Projekcja

Dysponujemy przykładami uczącymi opisanymi p-cechami $x=[x_1,x_2,\ldots,x_p]^T$ należącymi do dwóch klas C_1 *i* C_2

Wektory p-wymiarowe x są zrzutowane na prostą (kierunek związany z parametrami w). Algebraicznie odpowiada to zastąpieniu ich skalarem $z = w^T \cdot x$. Celem jest taki dobór w aby na podstawie nowej zmiennej z przykłady z obu klas były jak najlepiej rozdzielone.



Fisher LDA

Cel

- » Maksymalizuj odległość zrzutowanych średnich klas
- » Minimalizuj wariancje wewnątrz klasową

Odległość między rzutami średnich

$$(\mathbf{w}^T\bar{\mathbf{x}}_1 - \mathbf{w}^T\bar{\mathbf{x}}_2)^2$$

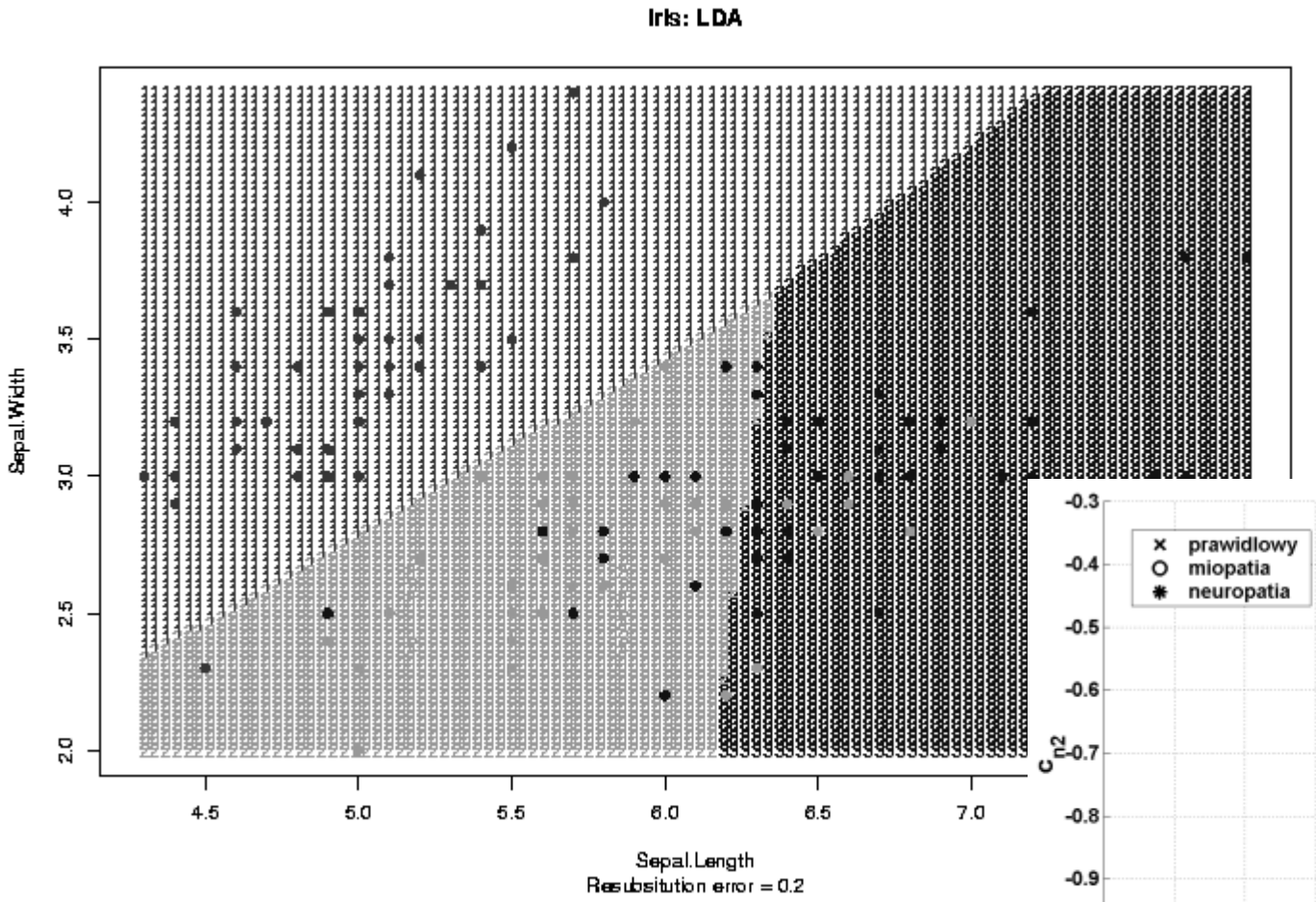
W celu maksymalizacji odległości rzutów średnich klas i minimalizacji wariancji wewnątrzklasowej należy poszukiwać wektora w który maksymalizuje następujące wyrażenie:

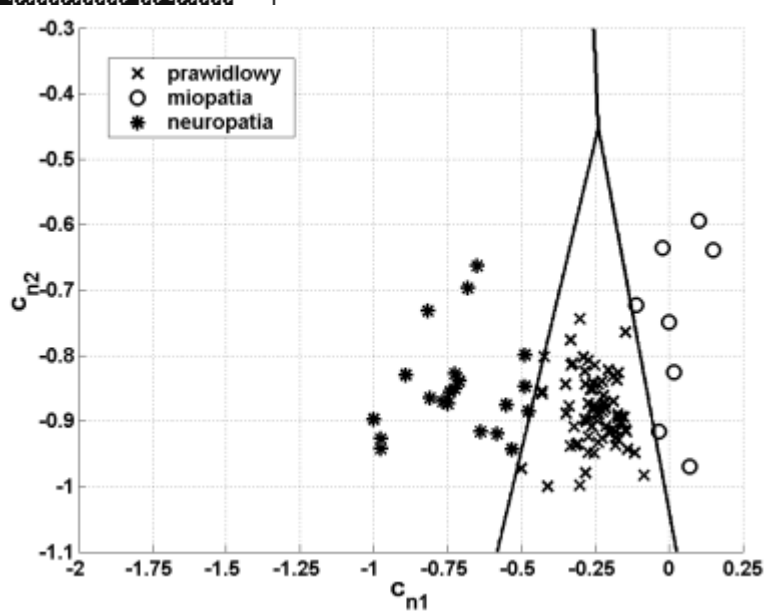
$$J(\mathbf{w}) = \frac{\left(\mathbf{w}^T\bar{\mathbf{x}}_1 - \mathbf{w}^T\bar{\mathbf{x}}_2\right)^2}{\mathbf{w}^T\mathbf{S}_W\mathbf{w}}$$

wskaźnik zmienności wewnątrzgrupowej

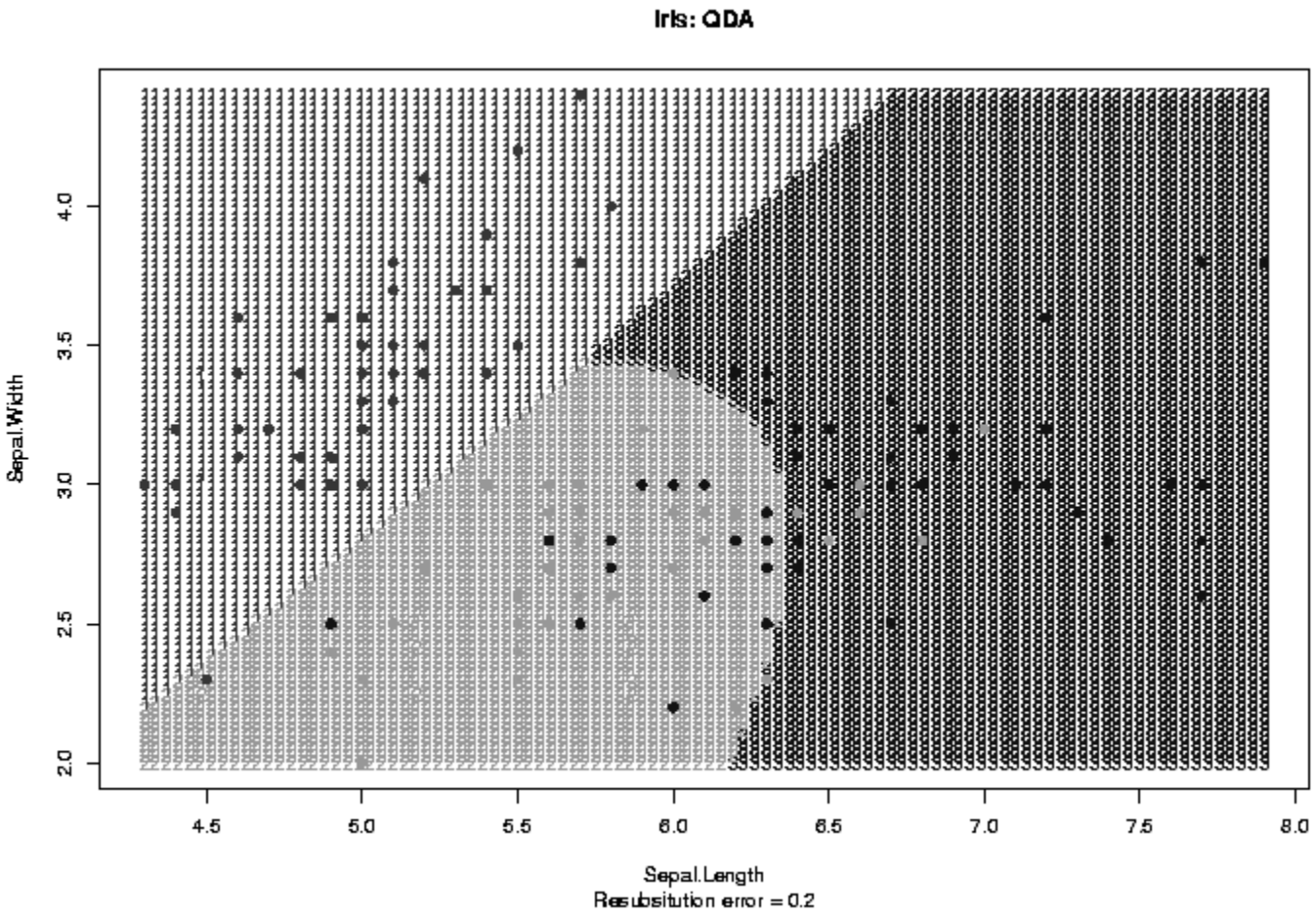
$$S_W = \frac{1}{n-2}\sum_{k=1}^{2}(n_k - 1)S_k$$

K=3

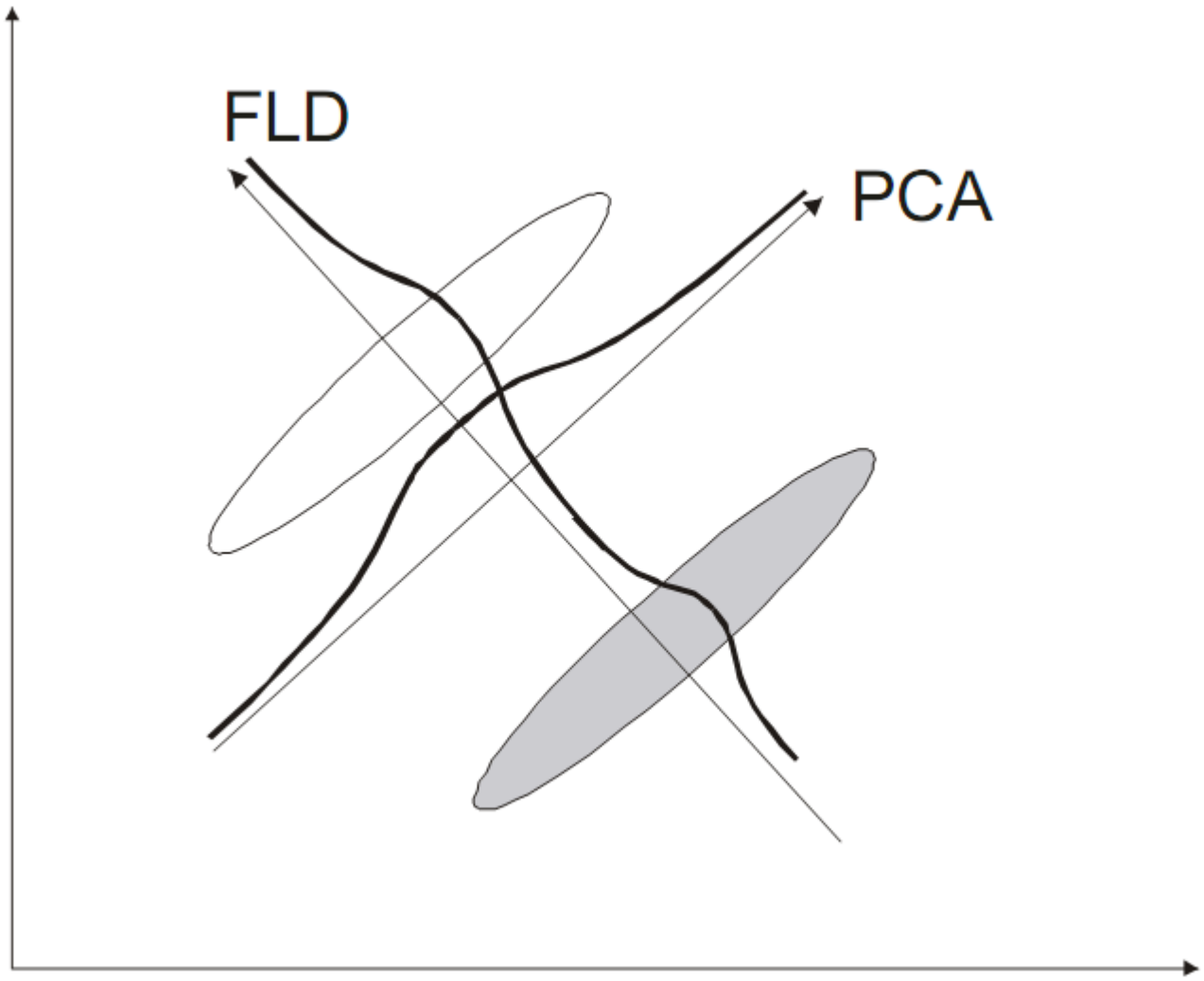

Iris: LDA
Sepal.Width
Sepal.Length
Resubsitution error = 0.2
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5
7.0


prawidlowy
miopatia
neuropatia
c_{n2}
c_{n1}
-0.3
-0.4
-0.5
-0.6
-0.7
-0.8
-0.9
-1
-1.1
-2
-1.75
-1.5
-1.25
-1
-0.75
-0.5
-0.25
0
0.25

QDA - *Quadratic Discriminant Analisys*



FLD a PCA



STATISTICA - przykład

dane: poziom rozwoju wybranych losowo 728 gmin w Polsce w 2005

cel: klasyfikacja gmin do grup: miejskich, wiejskich i miejsko-wiejskich na podstawie zmiennych o istotnej zdolności dyskryminacyjnej

wyjściowy zbiór zmiennych, które zostały następnie poddane ocenie zdolności dyskryminacyjnej, zawierał następujące zmienne:

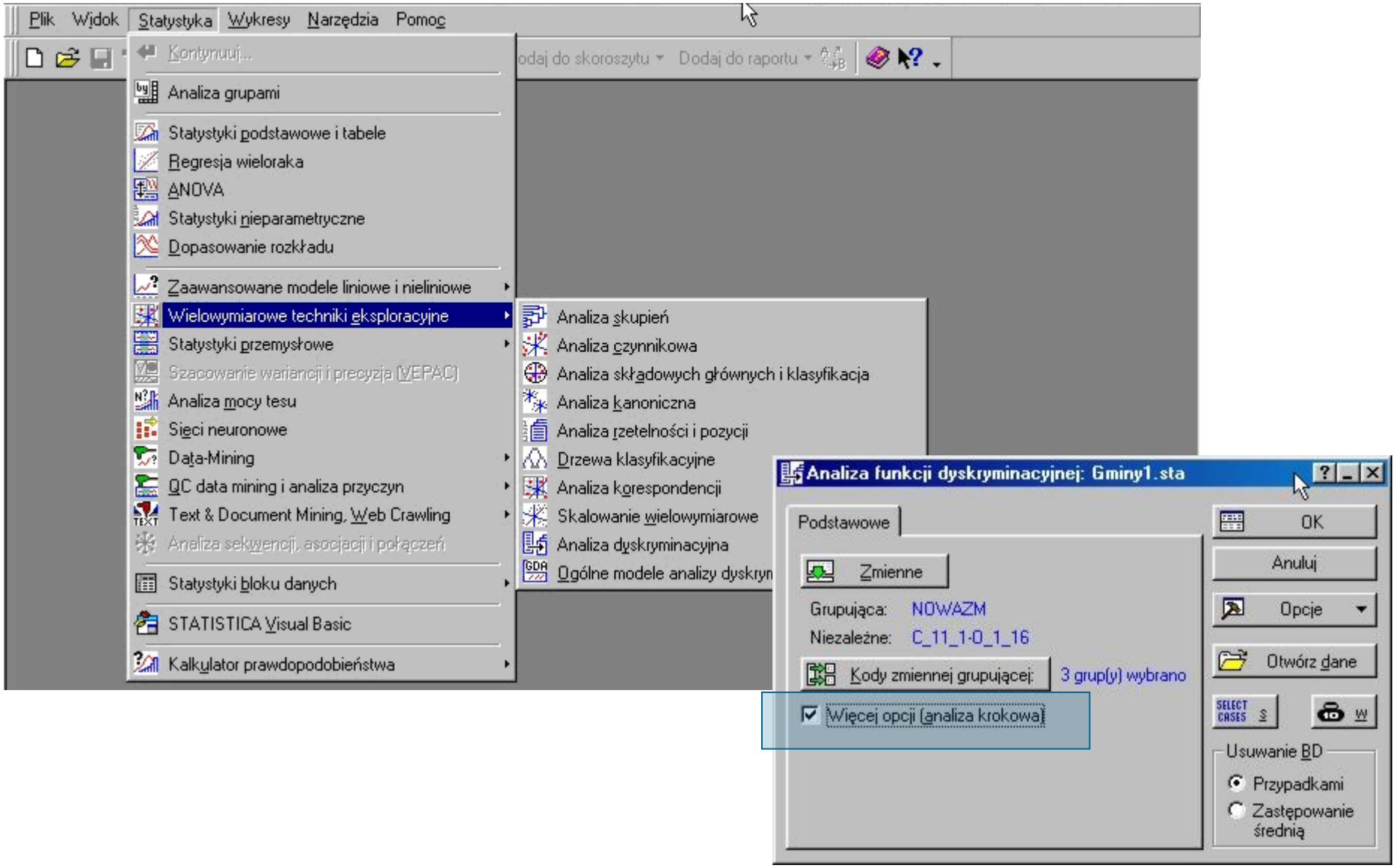
- C11.1 – liczba mieszkań ogółem na 1 mieszkańca,
- D1.1 – liczba aptek ogółem na 1 mieszkańca,
- E13.1 – liczba gimnazjów dla dzieci i młodzieży na 100 osób w wieku 13-15 lat,
- K1.1 – udział powierzchni użytków rolnych w powierzchni gminy ogółem,
- N1.1 – liczba jednostek (firm) zarejestrowanych w systemie REGON,
- O1.1 – dochody gminy ogółem w tys. zł na osobę,
- O1.1A – dochody własne gminy w tys. zł na osobę,
- O1.10 – subwencje ogólne w tys. zł na osobę,
- O1.12 – dotacje celowe z budżetu państwa w tys. zł na osobę,
- O1.16 – dotacje otrzymane z funduszy celowych w tys. zł na osobę.

Dane: Gminy1.sta (12 zmn. * 728 prz.)

	1 GMINA	2 NOWAZM	3 C 11 1	4 D 1 1	5 E 13 1	6 K 1 1	7 N 1 1	8 O 1 1	9 O 1 1A	10 O 1 10	C
1	02 01 01	1	1,134	1,123	-0,000	-1,201	1,884	0,157	0,473	-1,207	
2	02 01 02	1	1,563	0,028	-0,008	-0,294	0,749	-0,696	-0,112	-0,853	
3	02 01 02	1	1,679	0,642	-0,408	0,814	1,894	-0,499	0,309	-1,234	
4	02 01 02	1	0,821	-1,004	-0,606	-0,330	-0,389	-1,118	-0,629	-0,983	
5	02 01 02	1	0,521	0,293	-0,533	1,361	-0,480	-0,326	-0,400	-0,369	
6	02 01 05	1	0,712	1,087	-0,962	0,611	0,710	-0,272	0,350	-0,946	
7	02 01 06	1	2,517	0,990	3,021	-2,260	-0,250	4,220	3,699	-1,321	
8	02 01 06	1	1,137	0,586	-0,815	-1,693	-0,040	-0,790	-0,170	-1,182	
9	02 01 06	1	0,876	0,383	0,064	-1,676	-0,386	-0,619	0,294	-1,508	
10	02 01 06	1	1,727	0,207	1,375	-2,499	-0,191	1,090	1,845	-1,400	

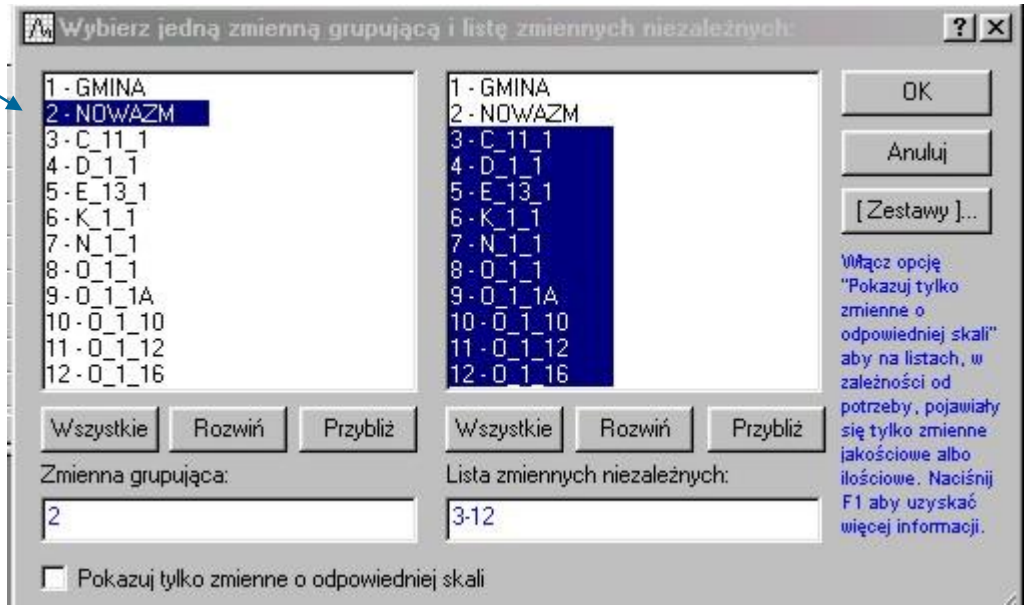
Fragment tablicy z danymi

moduł: Wielowymiarowe techniki eksploracyjne

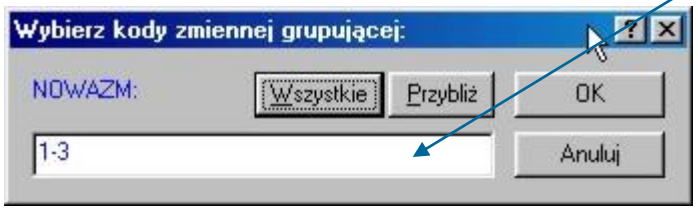


zmienne

zmienna grupująca



kody zmiennej grupującej: Wszystkie



1 - gmina miejska,

2 - gmina wiejska,

3 - gmina miejsko-wiejska.

okno: Definicja modelu

– procedura włączania do modelu/usuwania z modelu zmiennych zostaje zakończona gdy są spełnione pewne założenia zatrzymania procedury przez użytkownika.

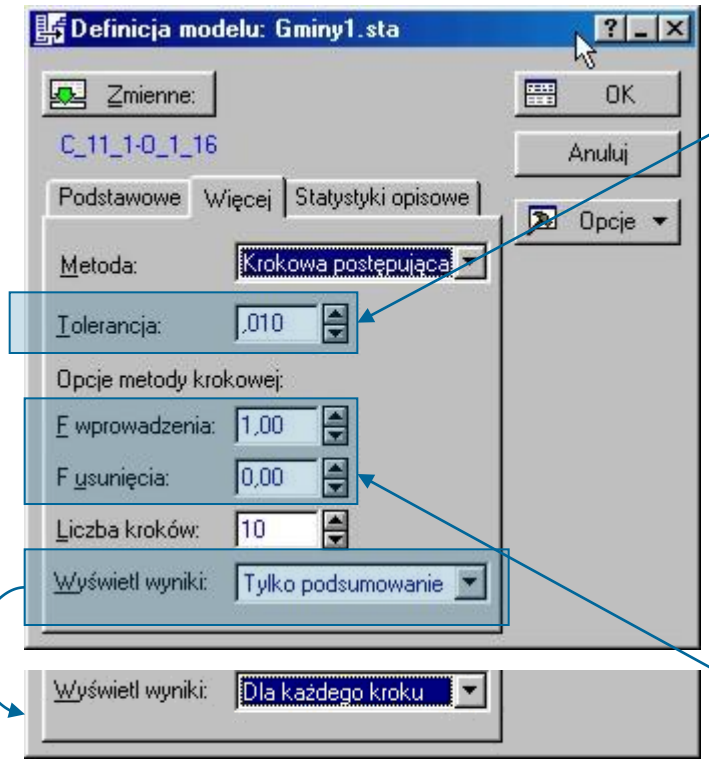
Metoda:

» **Standardowa** - wprowadzenie do modelu (równania funkcji dyskryminacyjnej) wszystkich wybranych zmiennych.

» **Krokowa postępująca** prowadzi do wprowadzania do modelu kolejnych zmiennych o najwyższej mocy dyskryminacyjnej.

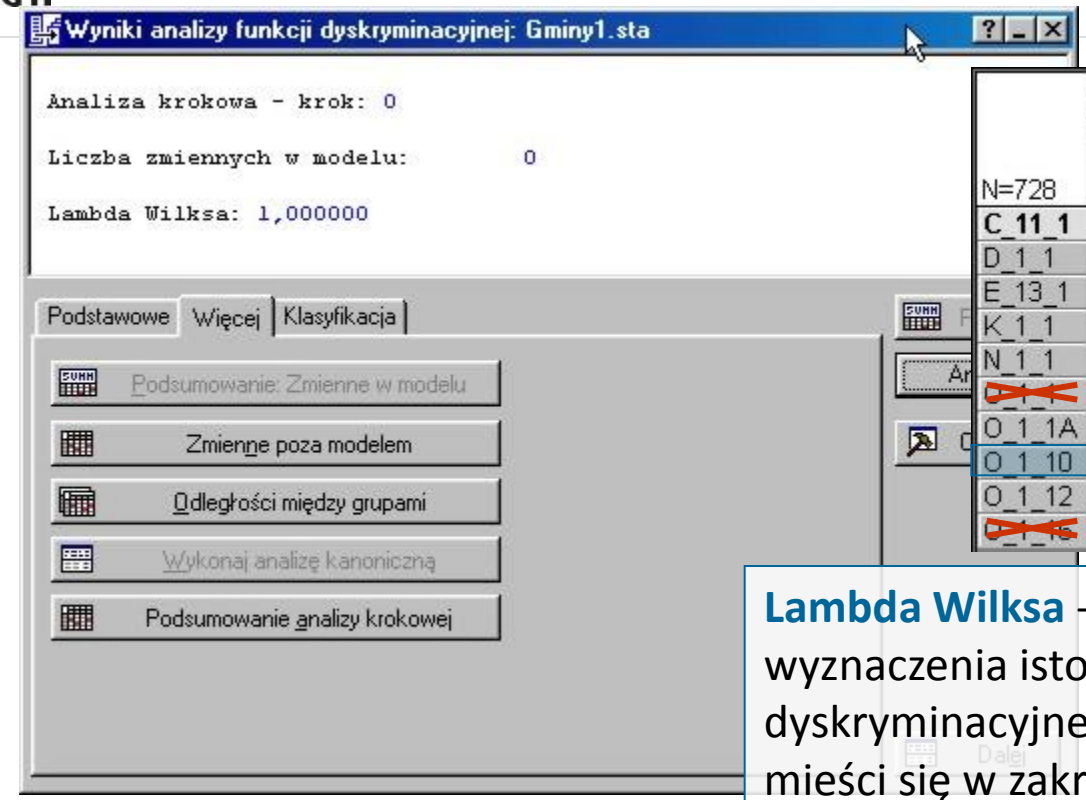
» **Krokowa wsteczna** powoduje wprowadzenie na początku do modelu wszystkich zmiennych, a następnie usuwanie z niego w kolejnych krokach zmiennych o najmniejszej mocy dyskryminacyjnej.

metoda krokowa postępująca



Tolerancja - jaki odsetek nowych informacji o gminach, nie powielanych ze zmiennymi już wprowadzonymi do modelu, musi wnosić dana zmienna aby została wprowadzona do modelu. Wartość 0,01 oznacza że nowa zmienna, aby zostać wprowadzona do modelu, musi wnosić do niego przynajmniej 1% nowych, nie wniesionych już do modelu przez znajdujące się w nim zmienne, informacji o badanych gminach

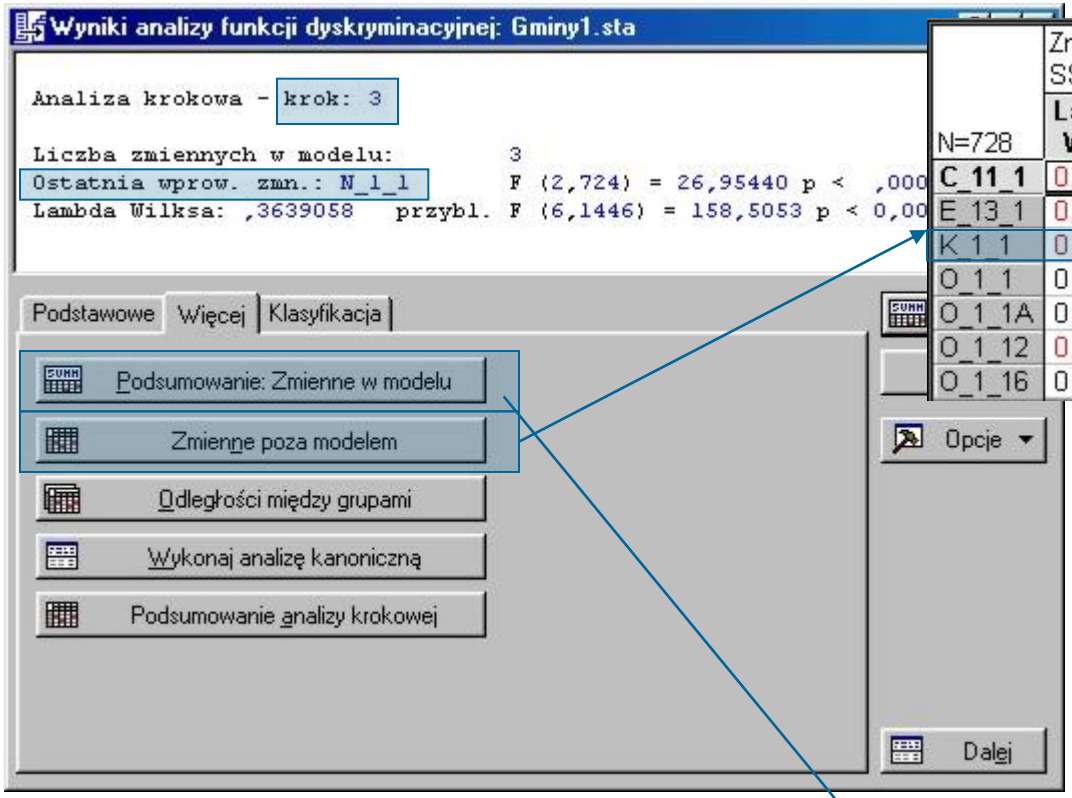
F wprowadzenia. Czym wyższa wartość tego parametru dla danej zmiennej, tym wyższa jej moc dyskryminacyjna. Jeżeli wartość parametru F dla danej zmiennej będzie większa, zmienna ta zostanie wprowadzona do modelu.



Zmienne aktualnie poza modelem (Gminy1.sta)
SS dla wszystkich F 2,725

N=728	Lambda Wilksa	Cząstk. Wilksa	F wprow.	poziom p	Toler.	1-Toler. (R-kwad)
C_11_1	0,657141	0,657141	189,1317	0,000000	1,000000	0,00
D_1_1	0,610194	0,610194	231,5732	0,000000	1,000000	0,00
E_13_1	0,911029	0,911029	35,4019	0,000000	1,000000	0,00
K_1_1	0,809498	0,809498	85,3083	0,000000	1,000000	0,00
N_1_1	0,662905	0,662905	184,3353	0,000000	1,000000	0,00
[illegible]	0,998636	0,998636	0,4950	0,609781	1,000000	0,00
O_1_1A	0,860510	0,860510	58,7620	0,000000	1,000000	0,00
O_1_10	0,566929	0,566929	276,9098	0,000000	1,000000	0,00
O_1_12	0,956974	0,956974	16,2980	0,000000	1,000000	0,00
[illegible]	0,999588	0,999588	0,1496	0,861115	1,000000	0,00

Lambda Wilksa - statystyka stosowana do wyznaczenia istotności statystycznej mocy dyskryminacyjnej aktualnego modelu. Jej wartość mieści się w zakresie od 1 (brak mocy dyskryminacyjnej) do 0 (maksymalna moc dyskryminacyjna). Uważajmy więc z interprertacją, bo mamy do czynienia z sytuacją odwrotną niż w przypadku większości poznanych już współczynników. Każda wartość podana w pierwszej kolumnie oznacza Lambdę Wilksa po wprowadzeniu tej zmiennej do modelu.



Zmienne aktualnie poza modelem (Gminy1.sta)
SS dla wszystkich F 2,722

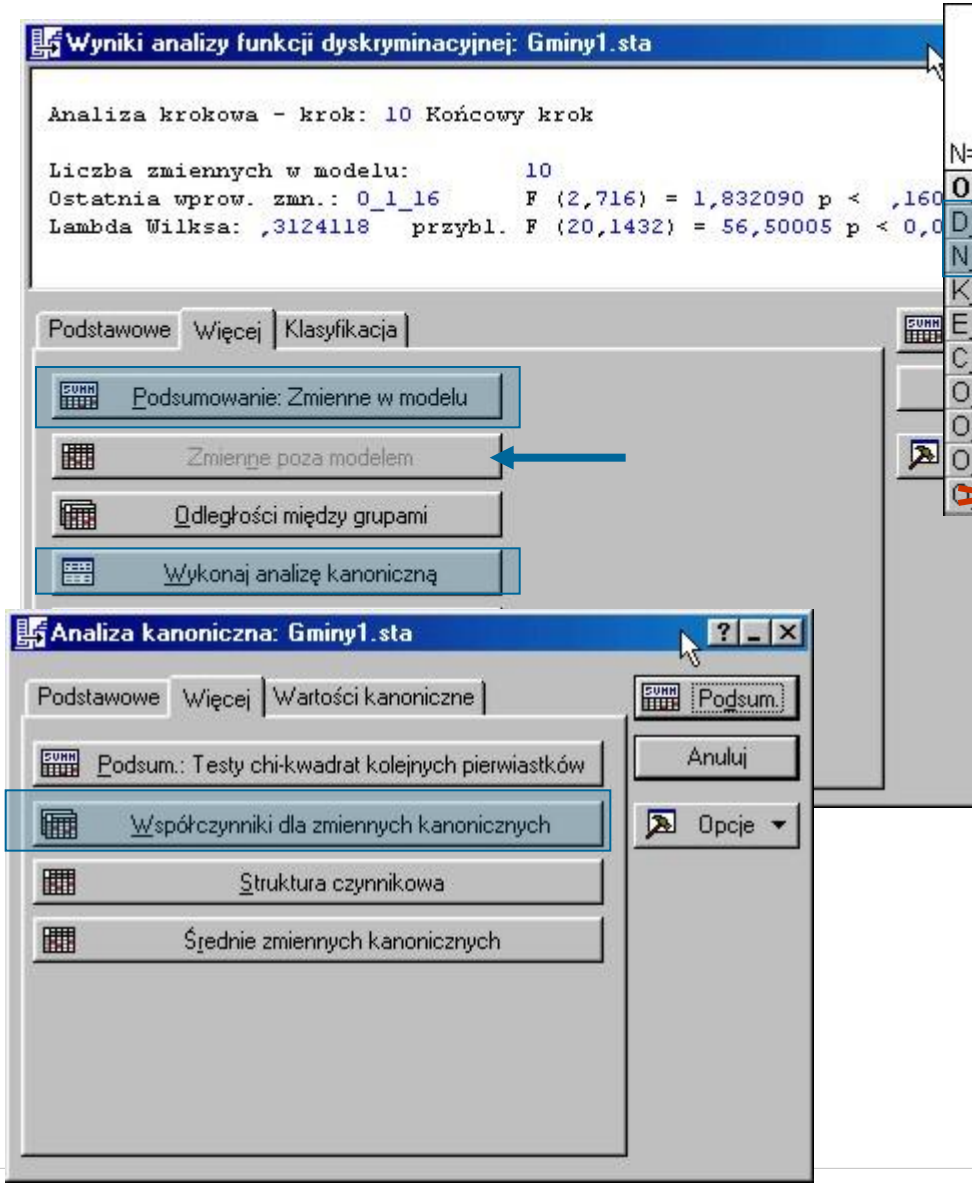
N=728	Lambda Wilksa	Cząstk. Wilksa	F wprow.	poziom p	Toler.	1-Toler. (R-kwad)
C_11_1	0,345183	0,948550	19,58099	0,000000	0,869919	0,130081
E_13_1	0,351740	0,966569	12,48617	0,000005	0,923148	0,076852
K_1_1	0,344419	0,946450	20,42523	0,000000	0,965351	0,034649
O_1_1	0,363338	0,998440	0,56420	0,569064	0,972209	0,027791
O_1_1A	0,363497	0,998878	0,40552	0,666783	0,855504	0,144496
O_1_12	0,358341	0,984708	5,60625	0,003837	0,857271	0,142729
O_1_16	0,363905	0,999998	0,00067	0,999333	0,999302	0,000698

82,6% informacji wnoszonych przez tą zmienną nie jest powielanych przez dwie pozostałe zmienne już znajdujące się w modelu

Podsumowanie analizy funkcji dyskryminacyjnej. (Gminy1.sta)
Krok 3, N zmn. w modelu: 3;Grupująca: NOWAZM (3 grup)
Lambda Wilksa: ,36391 przybl. F (6,1446)=158,51 p<0,0000

N=728	Lambda Wilksa	Cząstk. Wilksa	F usun. (2,723)	poziom p	Toler.	1-Toler. (R-kwad)
O_1_10	0,447832	0,812594	83,3716	0,000000	0,821628	0,178372
D_1_1	0,519045	0,701106	154,1140	0,000000	0,989324	0,010676
N_1_1	0,391040	0,930611	26,9544	0,000000	0,825936	0,174065

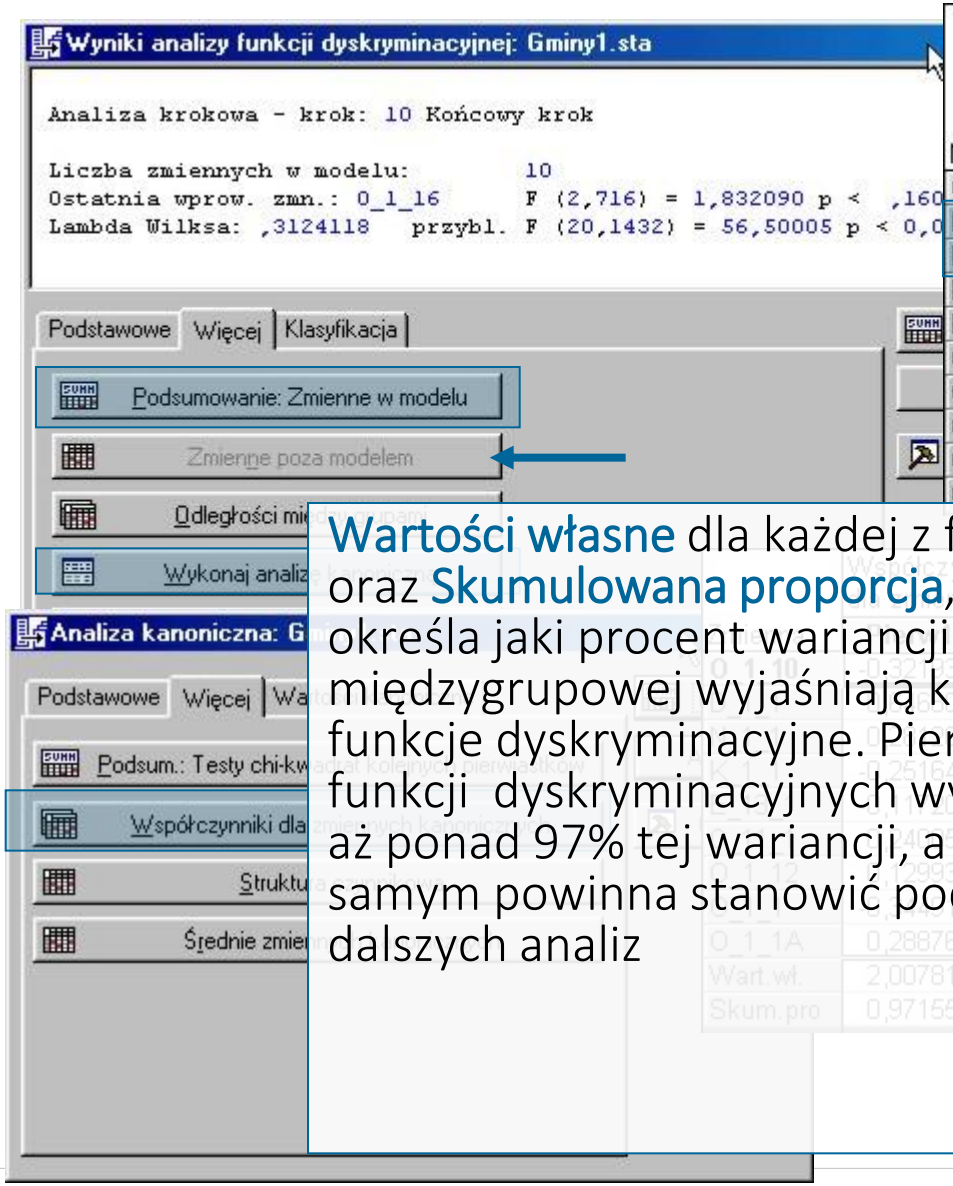
Wyniki analizy funkcji dyskryminacyjnej w kroku 10 końcowym



Podsumowanie analizy funkcji dyskryminacyjnej. (Gminy1.sta)
Krok 10, N zmn. w modelu: 10;Grupująca: NOWAZM (3 grup)
Lambda Wilksa: ,31241 przybl. F (20,1432)=56,500 p<0,0000

N=728	Lambda Wilksa	Cząstk. Wilksa	F usun. (2,716)	poziom p	Toler.	1-Toler. (R-kwad)
O_1_10	0,316771	0,986239	4,9951	0,007009	0,271694	0,728306
D_1_1	0,427167	0,731357	131,5011	0,000000	0,950390	0,049610
N_1_1	0,329669	0,947652	19,7759	0,000000	0,792965	0,207035
K_1_1	0,324903	0,961553	14,3143	0,000001	0,855238	0,144762
E_13_1	0,322349	0,969171	11,3877	0,000014	0,819424	0,180576
C_11_1	0,323015	0,967176	12,1500	0,000006	0,758709	0,241291
O_1_12	0,318240	0,981688	6,6781	0,001338	0,626347	0,373653
O_1_1	0,318847	0,979816	7,3745	0,000676	0,074371	0,925629
O_1_1A	0,317881	0,982795	6,2671	0,002004	0,075375	0,924625
~~O_1_15~~	0,314011	0,994908	1,8321	0,160830	0,635065	0,364935

Wyniki analizy funkcji dyskryminacyjnej w kroku 10 końcowym



Podsumowanie analizy funkcji dyskryminacyjnej. (Gminy1.sta)
Krok 10, N zmn. w modelu: 10;Grupująca: NOWAZM (3 grup)
Lambda Wilksa: ,31241 przybl. F (20,1432)=56,500 p<0,0000

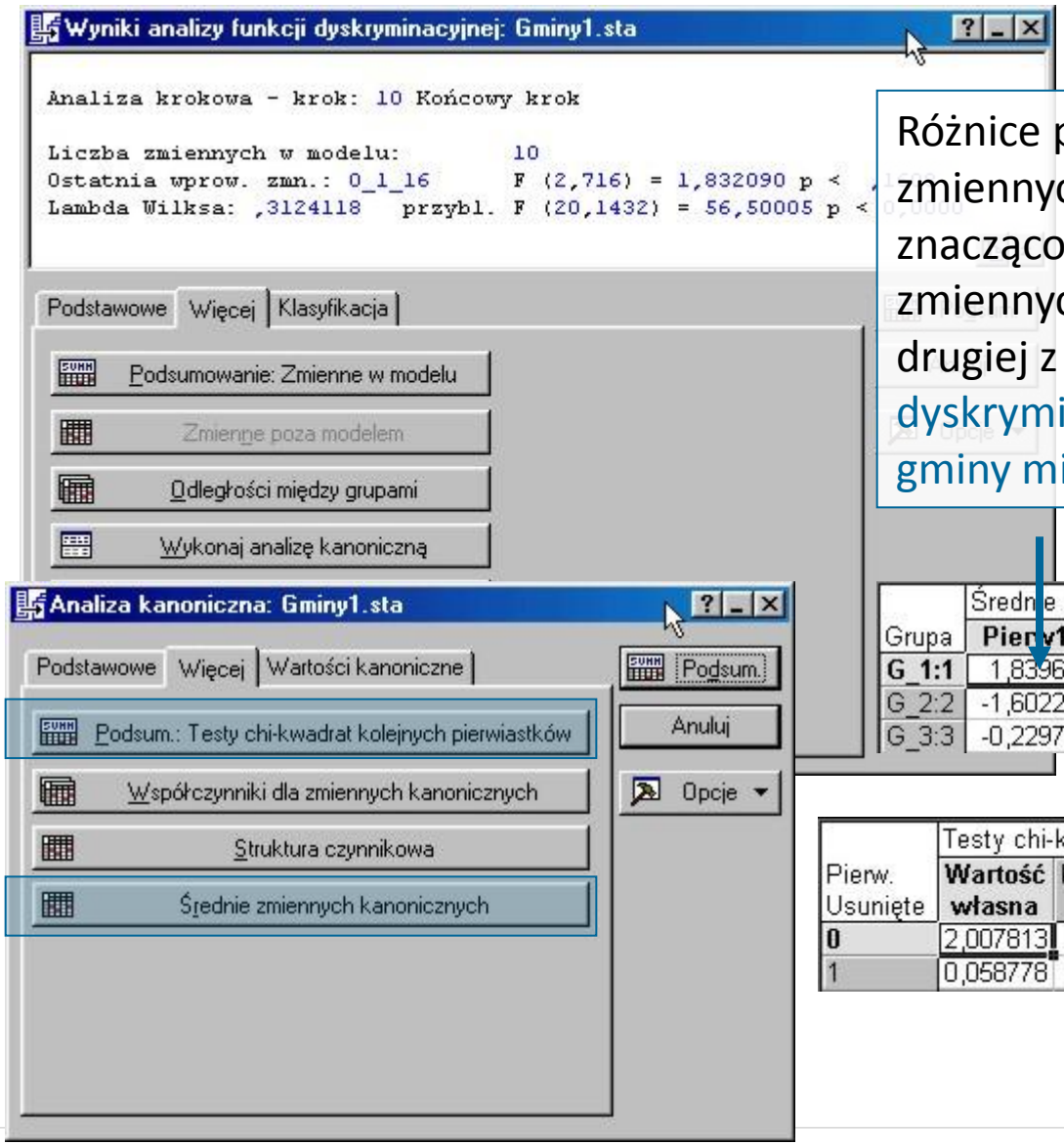
N=728	Lambda Wilksa	Cząstk. Wilksa	F usun. (2,716)	poziom p	Toler.	1-Toler. (R-kwad)
O_1_10	0,316771	0,986239	4,9951	0,007009	0,271694	0,728306
D_1_1	0,427167	0,731357	131,5011	0,000000	0,950390	0,049610
N_1_1	0,329669	0,947652	19,7759	0,000000	0,792965	0,207035
K_1_1	0,324903	0,961553	14,3143	0,000001	0,855238	0,144762
E_13_1	0,322349	0,969171	11,3877	0,000014	0,819424	0,180576
C_11_1	0,323015	0,967176	12,1500	0,000006	0,758709	0,241291
O_1_12	0,318240	0,981688	6,6781	0,001338	0,626347	0,373653
O_1_1	0,318847	0,979816	7,3745	0,000676	0,074371	0,925629
O_1_1A	0,317881	0,982795	6,2671	0,002004	0,075375	0,924625
[illegible]	0,314011	0,994908	1,8321	0,160830	0,635065	0,364935

Wartości własne dla każdej z funkcji oraz **Skumulowana proporcja**, która określa jaki procent wariancji międzygrupowej wyjaśniają kolejne funkcje dyskryminacyjne. Pierwsza z funkcji dyskryminacyjnych wyjaśnia aż ponad 97% tej wariancji, a tym samym powinna stanowić podstawę dalszych analiz

Surowe wsp. (Gminy1.sta) dla zmiennych kanonicz.

Zmienna	Pierw1	Pierw2
O_1_10	-0,426971	0,07289
D_1_1	0,826879	0,09413
N_1_1	0,344900	-0,61491
K_1_1	-0,279306	0,07898
E_13_1	-0,122630	-0,77191
C_11_1	0,295722	-0,35905
O_1_12	0,132645	0,47268
O_1_1	-0,344674	-1,06722
O_1_1A	0,310862	1,27329
Stała	-0,000000	-0,00000
Wart.wł.	2,007813	0,05878
Skum.pro	0,971558	1,00000

Wyniki analizy funkcji dyskryminacyjnej w kroku 10 końcowym

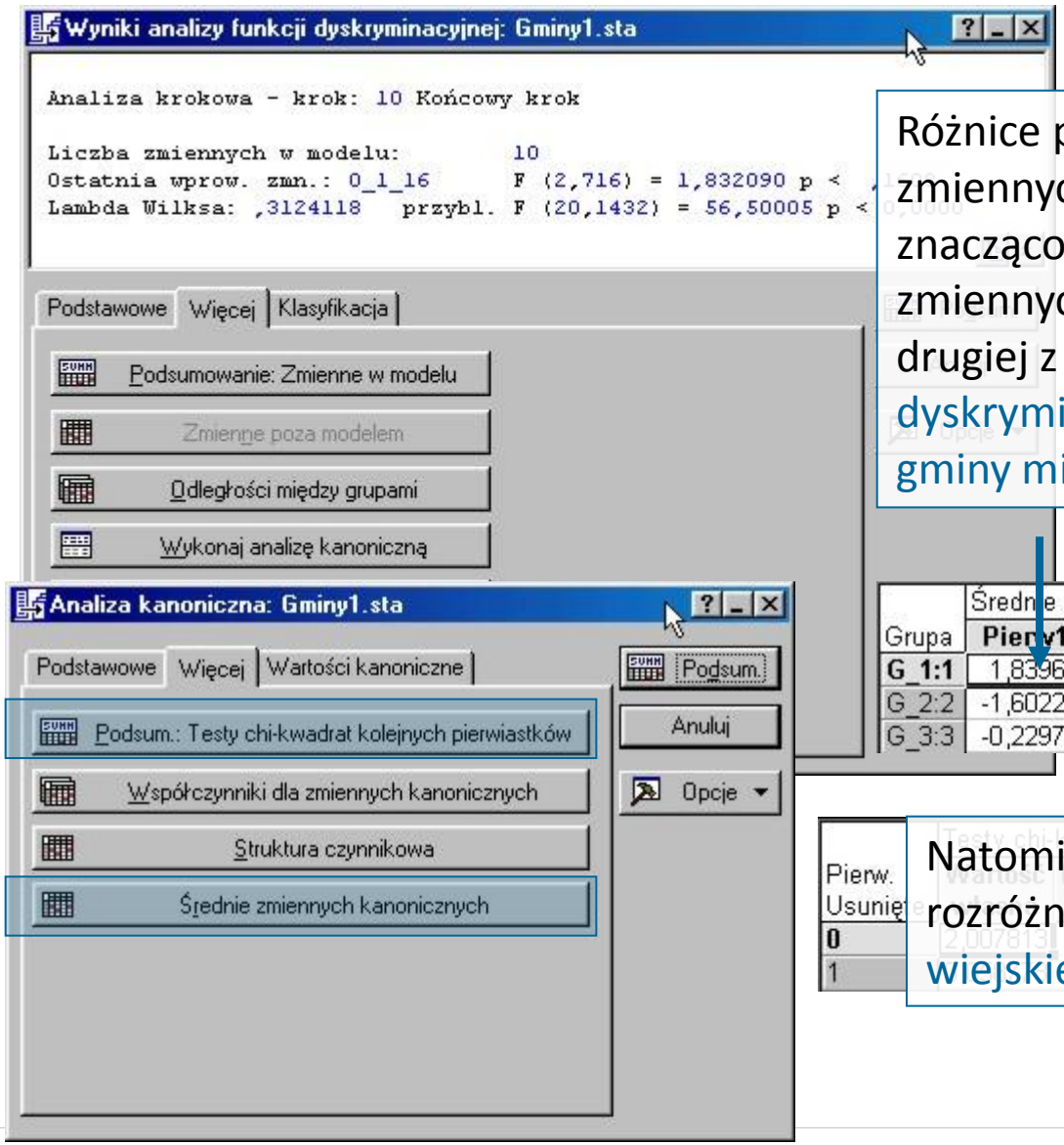


Różnice pomiędzy średnimi wartościami zmiennych dyskryminacyjnych dla gmin są znacząco większe dla pierwszej ze zmiennych dyskryminacyjnych niż dla drugiej z nich. Pierwsza funkcja dyskryminacyjna odróżnia przede wszystkim gminy miejskie od gmin wiejskich.

	Średnie zmien. kanonicznych	
Grupa	**Pierw1**	Pierw2
G_1:1	1,83962	-0,135958
G_2:2	-1,60229	-0,204139
G_3:3	-0,22975	0,339537

	Testy chi-kwadrat kolejnych pierwiastków (Gminy1.sta)					
Pierw. Usunięte	**Wartość własna**	Kanonicz R	Lambda Wilksa	chi-kwad	df	poziom p
0	2,007813	0,817027	0,314011	835,1550	18	0,000000
1	0,058778	0,235616	0,944485	41,1802	8	0,000002

Wyniki analizy funkcji dyskryminacyjnej w kroku 10 końcowym

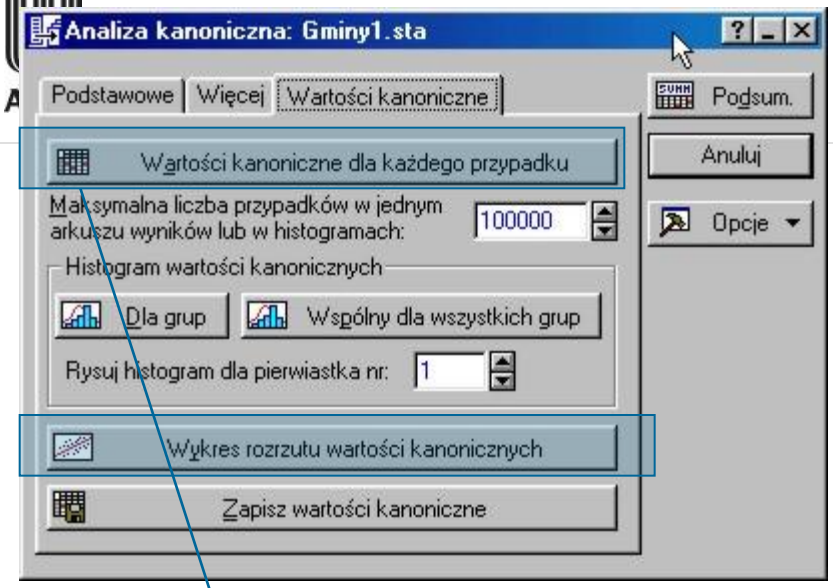


Różnice pomiędzy średnimi wartościami zmiennych dyskryminacyjnych dla gmin są znacząco większe dla pierwszej ze zmiennych dyskryminacyjnych niż dla drugiej z nich. Pierwsza funkcja dyskryminacyjna odróżnia przede wszystkim gminy miejskie od gmin wiejskich.

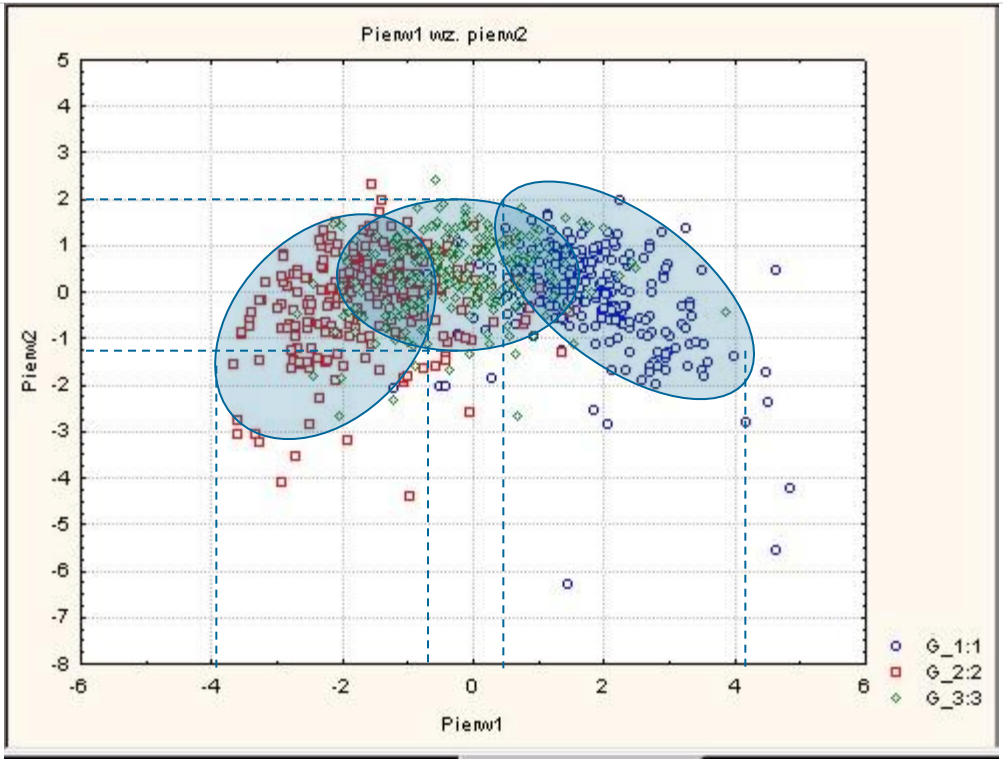
Grupa	Średnie zmien. kanonicznych Pierw1	Pierw2
G_1:1	1,83962	-0,135958
G_2:2	-1,60229	-0,204139
G_3:3	-0,22975	0,339537

Pierw. Usunię
0
1

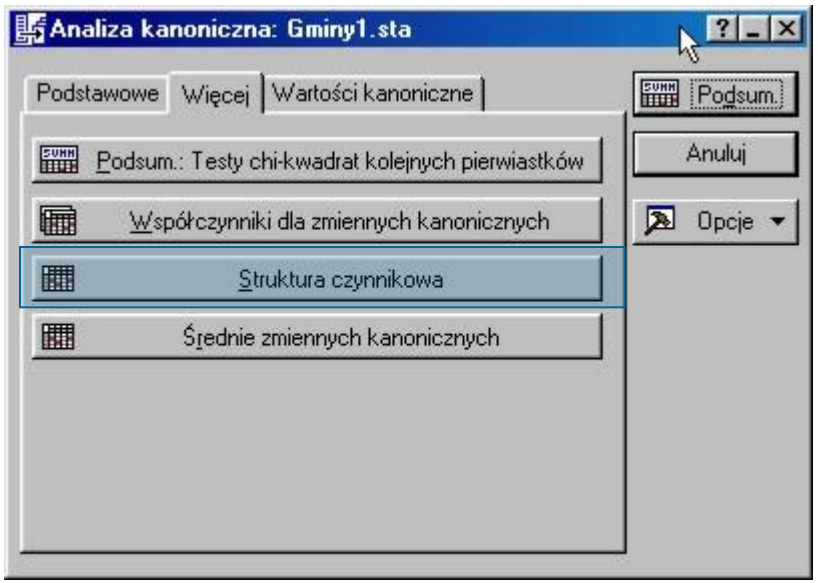
Natomiast druga funkcja dyskryminacyjna rozróżnia przede wszystkim gminy miejsko-wiejskie od pozostałych typów gmin.



Przyp.	Niestandaryzowane wart. kan		
	Grupa	Pierw1	Pierw2
1	G_1:1	2,76898	-1,52180
2	G_1:1	1,37582	-0,57207
3	G_1:1	2,24754	-0,67085
4	G_1:1	0,12806	0,87490
5	G_1:1	0,15232	0,78799
6	G_1:1	1,89071	0,80553
7	G_1:1	2,05991	-2,83218
8	G_1:1	2,17523	0,95942
9	G_1:1	1,83078	0,62760
10	G_1:1	1,96854	-0,55972
11	G_1:1	1,25860	-0,00710
12	G_1:1	2,78932	-1,99768
13	G_1:1	2,33979	-1,67298
14	G_1:1	1,42339	0,78761
15	G_1:1	1,50722	-0,13477
16	G_1:1	1,38434	-0,50164
17	G_1:1	1,60452	-0,04490
18	G_1:1	1,78089	0,46541
19	G_1:1	3,59336	-1,50434



Struktura czynnikowa

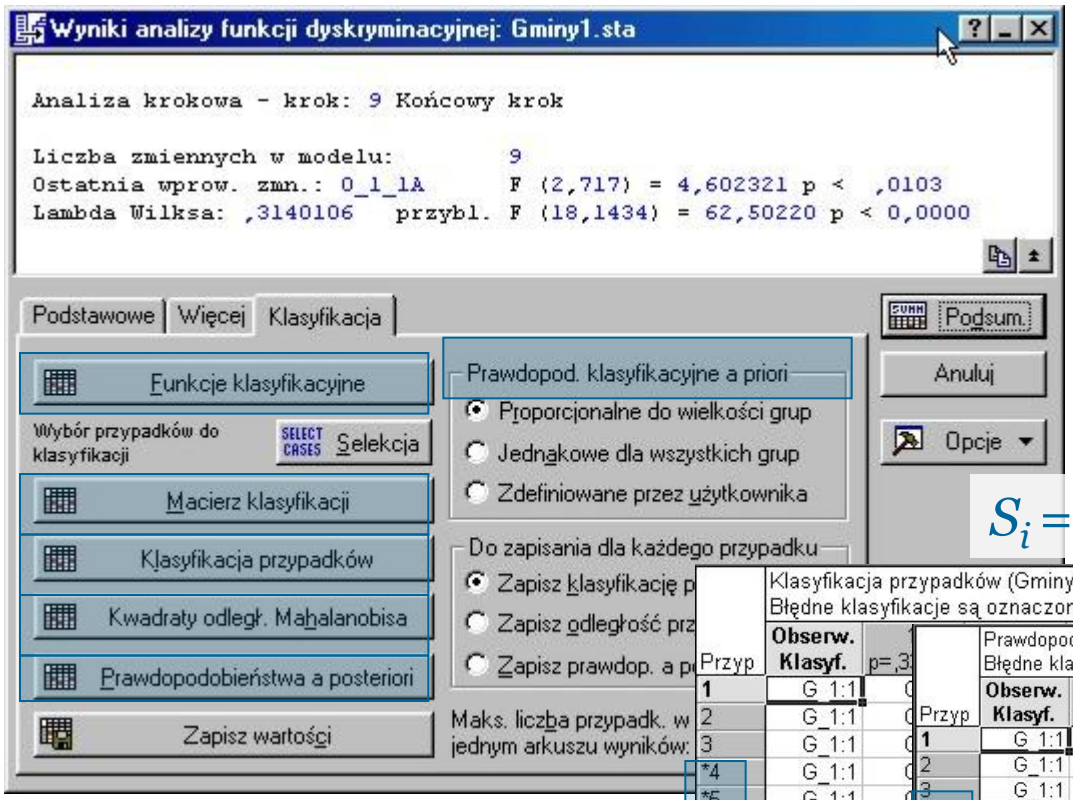


Macierz struk. czynnik. (Gminy1.sta)
Korelacje zmienne - pierwiastki kanor
(Zgrup. korelacje wewnątrzgrupowe)

Zmienna	Pierw1	Pierw2
O_1_10	-0,616199	-0,160852
D_1_1	0,564063	-0,006245
N_1_1	0,500988	-0,278891
K_1_1	-0,341241	0,161471
E_13_1	-0,182009	-0,727946
C_11_1	0,506354	-0,343885
O_1_12	-0,146025	0,191102
O_1_1	-0,011826	-0,135846
O_1_1A	0,284134	-0,010885

współczynniki korelacji – ładunki czynnikowe

Klasyfikacja



Zmienna	Funkcje klasyfikacyjne; grupując G_1:1 p=,33242	G_2:2 p=,33379	G_3:3 p=,33379
O_1_10	-0,79537	0,66925	0,12285
D_1_1	1,50834	-1,34412	-0,15802
N_1_1	0,71809	-0,42710	-0,28803
K_1_1	-0,52455	0,43141	0,09099
E_13_1	-0,12064	0,35407	-0,23392
C_11_1	0,59283	-0,40054	-0,18985
O_1_12	0,17975	-0,30903	0,13002
O_1_1	-0,48897	0,77013	-0,28317
O_1_1A	0,39875	-0,75802	0,36091
Stała	-2,80270	-2,40175	-1,18128

w_{i1}

$$S_i = c_i + w_{i1}*x_1 + w_{i2}*x_2 + \ldots + w_{im}*x_m$$

Macierz klasyfikacji (Gminy1.sta)
Wiersze: obserwowana klasyfik.
Kolumny: Przewidywana klasyfikacja

Grupa	Procent Poprawne	G_1:1 p=,33242	G_2:2 p=,33379	G_3:3 p=,33379
G_1:1	85,53719	207	3	32
G_2:2	75,72016	6	184	53
G_3:3	64,19753	29	58	156
Razem	75,13736	242	245	241

Klasyfikacja przypadków (Gminy1.sta)
Błędne klasyfikacje są oznaczone *

Przyp	Obserw. Klasyf.	p=,3
1	G_1:1	
2	G_1:1	
3	G_1:1	
*4	G_1:1	
*5	G_1:1	
6	G_1:1	
7	G_1:1	
8	G_1:1	
9	G_1:1	
10	G_1:1	
11	G_1:1	
12	G_1:1	
13	G_1:1	
14	G_1:1	
15	G_1:1	
16	G_1:1	
17	G_1:1	

Prawdopod. a posteriori (Gminy1.sta)
Błędne klasyfikacje są oznaczone *

Przyp	Obserw. Klasyf.	G_1:1 p=,33242	G_2:2 p=,33379	G_3:3 p=,33379
1	G_1:1	0,991976	0,000119	0,007905
2	G_1:1	0,808226	0,011016	0,180758
3	G_1:1	0,965421	0,000659	0,033920
*4	G_1:1	0,128362	0,116209	0,755429
*5	G_1:1	0,138733	0,116221	0,745046
6	G_1:1	0,869212	0,001833	0,128955
7	G_1:1	0,980347	0,001480	0,018173
8	G_1:1	0,917957	0,000719	0,081323
9	G_1:1	0,864344	0,002268	0,133389
10	G_1:1	0,936546	0,001658	0,061796
11	G_1:1	0,718117	0,014099	0,267784
12	G_1:1	0,993829	0,000115	0,006056
13	G_1:1	0,981780	0,000523	0,017697
14	G_1:1	0,715591	0,007547	0,276862
15	G_1:1	0,820881	0,006909	0,172211
16	G_1:1	0,806192	0,010620	0,183188
17	G_1:1	0,843881	0,005050	0,151069

Kwadraty odległości Mahalanobisa od cent
Błędne klasyfikacje są oznaczone *

	G_1:1 p=,33242	G_2:2 p=,33379	G_3:3 p=,33379
1	4,6516	22,7116	14,3243
1	3,2590	11,8583	6,2626
1	7,0773	21,6638	13,7826
1	10,9182	11,1254	7,3816
1	8,8599	9,2223	5,5063
1	2,6880	15,0195	6,5125
1	32,7426	45,7427	40,7268
1	6,4801	20,7911	11,3357
1	7,7793	19,6739	11,5249
1	11,3426	24,0238	16,7876
1	2,5353	10,4046	4,5164
1	21,2080	39,3432	31,4175
1	13,0563	28,1412	21,0965
1	3,7499	12,8620	5,6574
1	3,9529	13,5163	7,0844
1	7,7802	16,4476	10,7520
1	1,9016	12,1471	5,3504

Zbiory przybliżone

Zbiory przybliżone

Podobnie jak logika rozmyta, logika przybliżona zajmuje się modelowaniem niepewności.

Niepewność wynika z granularności informacji

Podstawowym zastosowaniem logiki przybliżonej jest klasyfikacja, logika ta pozwala na budowanie modeli aproksymacji zbiorów, do których przynależność jest określana na podstawie atrybutów.

Zbiory definiowane są atrybutami, nie jak w klasycznej teorii mnogości poprzez ich elementy.

Logika przybliżona rozwijana była jako jedna z metod eksploracji wiedzy (*data mining*) .

System Informacyjny

definicja systemu informacyjnego w oparciu o agregat:

$$S = \langle U, A, V, f : U \times A \rightarrow V \rangle$$

gdzie:

U – jest niepustym i skończonym zbiorem obiektów zwanym uniwersum;

A – jest zbiorem atrybutów;

V – jest dziedziną atrybutu a∈A;

$f : U \times A \rightarrow V$ jest funkcją informacyjną, taką że

$$\forall a \in A,\ x \in U,\ f(a, x) \in V_a$$

Tablica decyzyjna

Jeżeli w systemie informacyjnym wyróżniamy rozłączne zbiory atrybutów warunkowych C i atrybutów decyzyjnych D gdzie $A = C \cup D$, to system taki nazywany jest tablicą decyzyjną.

Symbol atrybutu	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9
Symbol obiektu w tablicy	**Symbol wady**	**Nazwa wady**	**Norma**	**rodzaj uszkodzenia**	**rozmieszczenie**	**lokalizacja**	**liczba uszkodzeń**	**kształt uszkodzeń**	**operacja technologiczna**
X_1	341	FAŁDY	CZ	zmarszczki, rysa, rozmycia	miejscowe	ścianka wkładka, podpórka rdzenia, powierzchnia	liczne	wąska, zaokrąglone brzegi	projektowanie odlewu, zalewanie, chłodzenie odlewu
X_2	W207	FAŁDA	PL	szczelina, rysa	miejscowe	powierzchnia	pojedyncze	wąska, zaokrąglone brzegi	projektowanie układu wlewowego, zalewanie,
X_3	W407	ZIMNE KROPLE	PL	krople		wnętrze		kuliste	projektowanie układu wlewowego, zalewanie
X_4	C311	FAŁDA, ZIMNA KROPLA	FR	przerwanie ciągłości, szczelina	rozległe	powierzchnia, pod powierzchnią	liczne	zaokrąglone brzegi, wąska	projektowanie układu zasilania, zalewanie
X_5	C331	FAŁDA W SĄSIEDZTWIE RDZENIA LUB INNEJ CZEŚCI METALOWEJ	FR	przerwanie ciągłości	miejscowe	w sąsiedztwie wkładek	brak danych	zakrzywione ścianki	zalewanie, krzepnięcie

Zbiory przybliżone

Elementy, o których mamy identyczną informację są nierozróżnialne i tworzą tzw. zbiory elementarne (granule).

O elementach znajdujących się w obszarze zbioru elementarnego możemy powiedzieć jedynie, że wszystkie wartości ich atrybutów są takie jak całego zbioru elementarnego.

Suma dowolnych zbiorów elementarnych jest nazywana zbiorem definiowalnym.

Zbiory, które nie są zbiorami definiowalnymi nazywane są zbiorami przybliżonymi.

Zbiory przybliżone

Zbiór przybliżony to para klasycznych zbiorów: przybliżenie dolne i przybliżenie górne

na zbiorach przybliżonych podstawowe działania są takie same, jak działania na zbiorach klasycznych. Dodatkowo wprowadza się kilka nowych pojęć, które nie są używane w przypadku zbiorów klasycznych.

dla każdego podzbioru cech $B \subseteq A$ pary obiektów pozostają w relacji nierozróżnialności jeśli posiadają takie same wartości dla wszystkich atrybutów ze zbioru B, co można zapisać:

(*indiscernibility relation*)

$$IND(B) = \{x_i, x_j \in U : \forall b \in B,\ f(x_i, b) = f(x_j, b)\}$$

Relacja nierozróżnialności

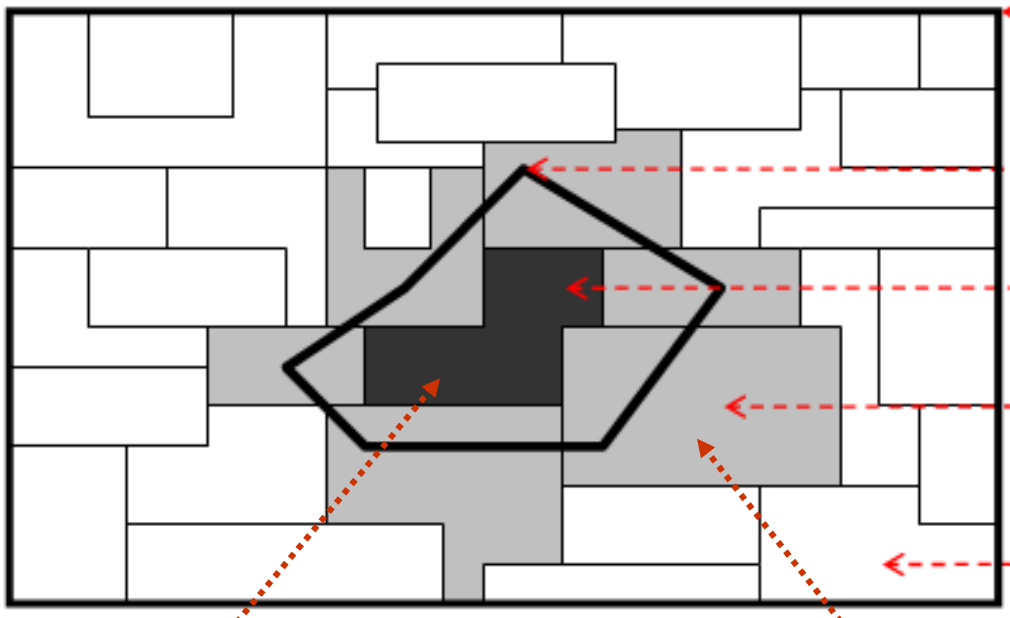
Każda relacja nierozróżnialności dzieli zbiór na rodzinę rozłącznych podzbiorów zwanych klasami abstrakcji (równoważności) lub zbiorami elementarnymi.

zbiór $[x_i]_{IND(B)}$ zawiera wszystkie obiekty systemu, które są nierozróżnialne z obiektem x_i po atrybutach B.

relacja nierozróżnialności opisuje zjawisko, że system informacyjny nie jest w stanie wskazać jako indywiduum obiektu spełniającego wartości podanych atrybutów w warunkach niepewności (nieokreśloności niektórych atrybutów nieuwzględnionych w systemie).

System zwraca zbiór wartości atrybutów pasujących do wskazanego obiektu będący pewną aproksymacją.

Aproksymacja



elementy bez wątpliwości należą do zbioru

$$\underline{S}(X) = \{x \in U : [x]_{IND(B)} \subseteq X\}$$

elementów nie można wykluczyć

$$\overline{S}(X) = \{x \in U : [x]_{IND(B)} \cap X \neq \emptyset\}$$

Dokładność aproksymacji określa wyrażenie:

$$\mu(a,U) = \frac{card\,\underline{S}}{card\,\overline{S}}$$

gdzie: *card* – symbol określający moc (liczbę elementów) danego zbioru.

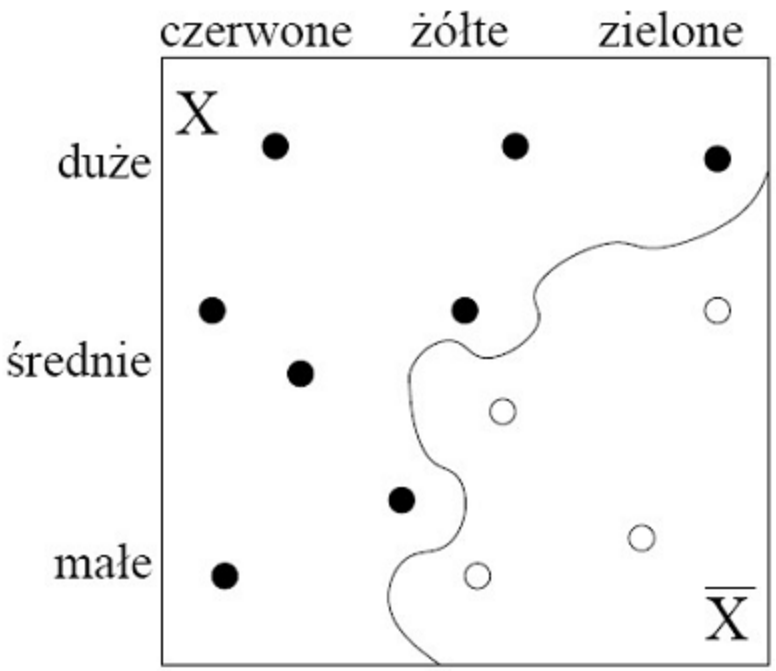
Aproksymacja

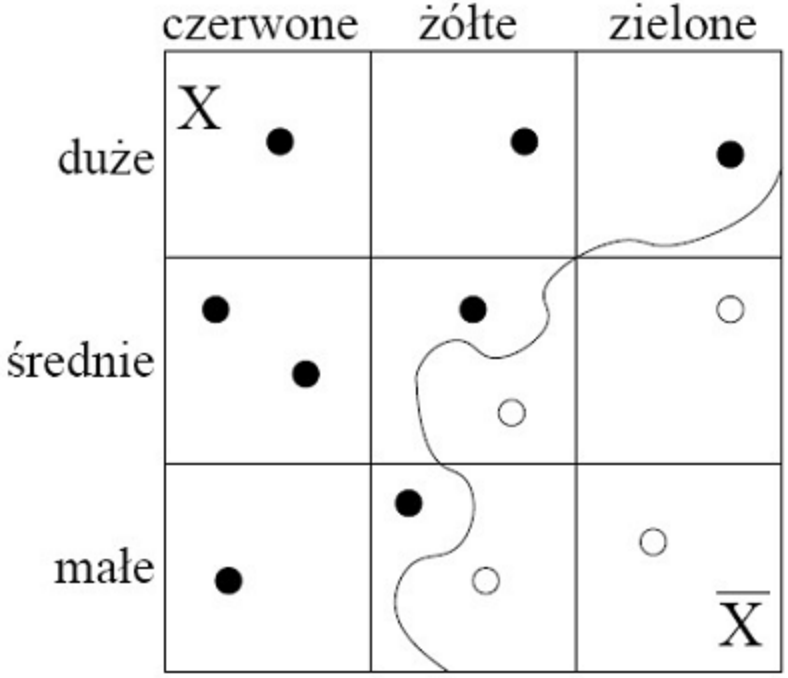
Za pomocą dolnej i górnej aproksymacji jesteśmy w stanie określić nieostre pojęcie w ścisły sposób.

Aproksymacja dolna oznacza, że elementy bez wątpliwości należą do zbioru (w świetle posiadanej wiedzy mogą być zaklasyfikowane jednoznacznie do rozważanego zbioru) .

Brzeg zawiera tylko te obiekty z górnego przybliżenia, które mogą być tylko uznane za możliwie należące do X, na podstawie atrybutów (nie można ich wykluczyć, w świetle posiadanej wiedzy, z danego zbioru), których nie można jednoznacznie przydzielić do X z uwagi na niepełny opis atrybutów.

Przykład klasyfikacji



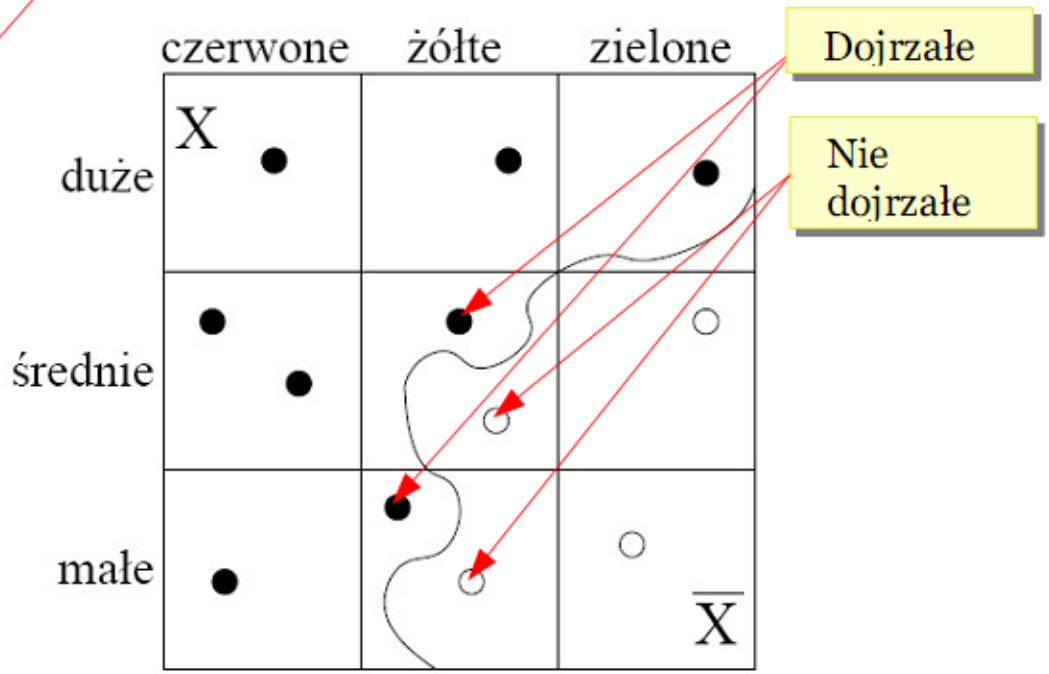


X – jabłka dojrzałe, $\overline{X}$ – jabłka niedojrzałe.

	kolor	*wielkość*	*dojrzałe*
x_1	*czerwone*	*duże*	*tak*
x_2	*żółte*	*średnie*	*tak*
x_3	*zielone*	*małe*	*nie*
x_4	*zielone*	*duże*	*tak*
x_5	*żółte*	*średnie*	*nie*
x_6	*czerwone*	*średnie*	*tak*
x_7	*żółte*	*duże*	*tak*
x_8	*czerwone*	*średnie*	*tak*
x_9	*żółte*	*małe*	*nie*
x_{10}	*żółte*	*małe*	*tak*
x_{11}	*czerwone*	*małe*	*tak*
x_{12}	*zielone*	*średnie*	*nie*

Gdy *żółte* i *średnie* to dojrzałe czy nie?

Gdy *żółte* i *małe* to dojrzałe czy nie?



Niech:

$B = \{ kolor, wielkość \}$

$X_n = \{ x_3, x_5, x_9, x_{12} \}$

	kolor	*wielkość*	*dojrzałe*
x_1	*czerwone*	*duże*	*tak*
x_2	*żółte*	*średnie*	*tak*
x_3	*zielone*	*małe*	*nie*
x_4	*zielone*	*duże*	*tak*
x_5	*żółte*	*średnie*	*nie*
x_6	*czerwone*	*średnie*	*tak*
x_7	*żółte*	*duże*	*tak*
x_8	*czerwone*	*średnie*	*tak*
x_9	*żółte*	*małe*	*nie*
x_{10}	*żółte*	*małe*	*tak*
x_{11}	*czerwone*	*małe*	*tak*
x_{12}	*zielone*	*średnie*	*nie*

$[x_1]_{IND(B)} = \{ x_1 \}$

$[x_2]_{IND(B)} = \{ x_2, x_5 \}$

$[x_3]_{IND(B)} = \{ x_3 \}$

$[x_4]_{IND(B)} = \{ x_4 \}$

$[x_5]_{IND(B)} = \{ x_2, x_5 \}$

$[x_6]_{IND(B)} = \{ x_6, x_8 \}$

$[x_7]_{IND(B)} = \{ x_7 \}$

$[x_8]_{IND(B)} = \{ x_6, x_8 \}$

$[x_9]_{IND(B)} = \{ x_9, x_{10} \}$

$[x_{10}]_{IND(B)} = \{ x_9, x_{10} \}$

$[x_{11}]_{IND(B)} = \{ x_{11} \}$

$[x_{12}]_{IND(B)} = \{ x_{12} \}$

Przybliżenie dolne:

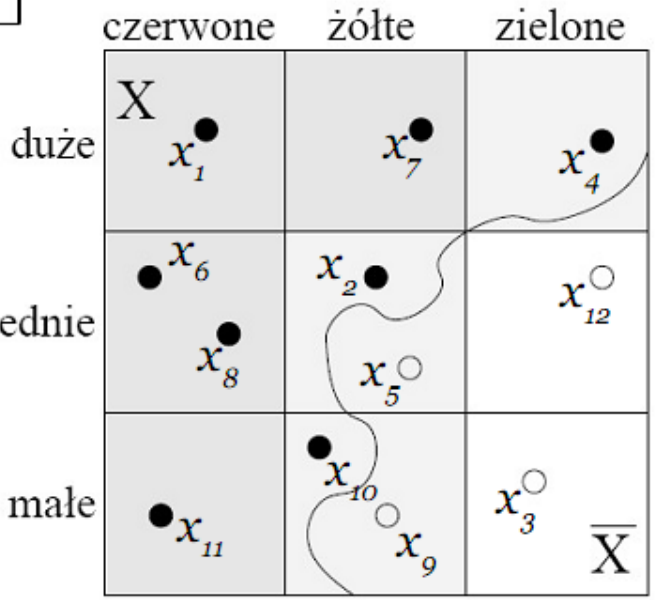
$B_{IND(B)}X_n = \{ x_3, x_{12} \}$

Przybliżenie górne:

$B^{IND(B)}X_n = \{ x_2, x_3, x_5, x_9, x_{10}, x_{12} \}$

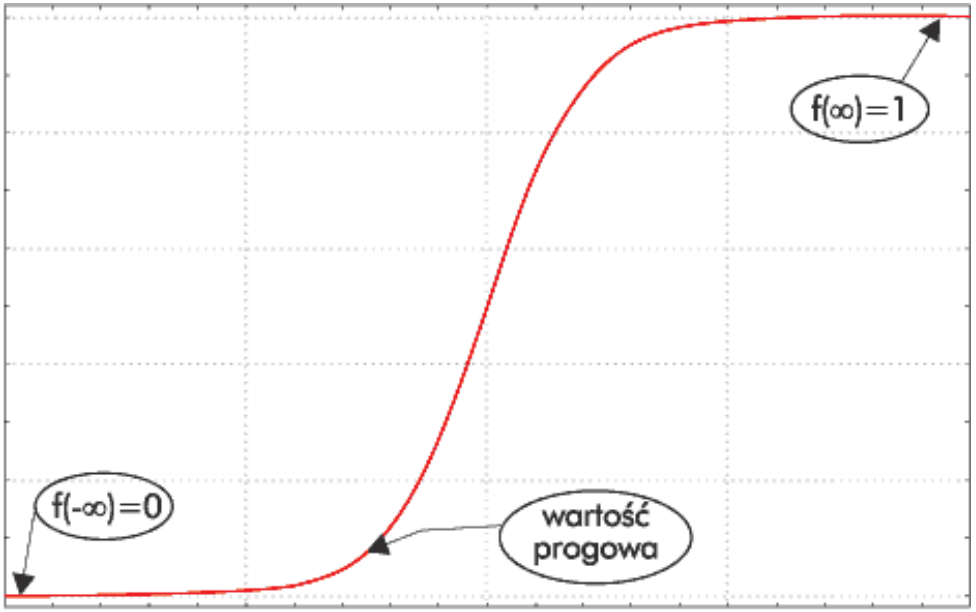
Brzeg:

$BN_N(X_n) = \{ x_2, x_5, x_9, x_{10} \}$



Logistyczne Funkcje Dyskryminacyjne

Regresja logistyczna



» estymacja modelu, który opisuje zależność między jedną lub kilkoma ciągłymi zmiennymi niezależnymi a binarną zmienną zależną.

Regresja logistyczna

Modele dla odpowiedzi binarnych:

- Na przykład pacjenci powrócą do zdrowia po urazie albo nie; kandydaci do pracy przejdą albo nie przejdą testu kwalifikacyjnego, kupony mogą zostać lub nie zostać zwrócone itd.
- można zastosować procedury standardowej regresji wielorakiej i obliczyć standardowe współczynniki regresji.
- model prowadzi do przewidywanych wartości większych niż 1 lub mniejszych niż 0. Jednakże przewidywane wartości, które są większe niż 1 lub mniejsze niż 0 nie są prawidłowe; tak więc, gdy stosuje się standardową procedurę regresji wielorakiej, ograniczenie zakresu zmiennej binarnej (np. między 0 a 1) jest ignorowane.

Regresja logistyczna (logit)

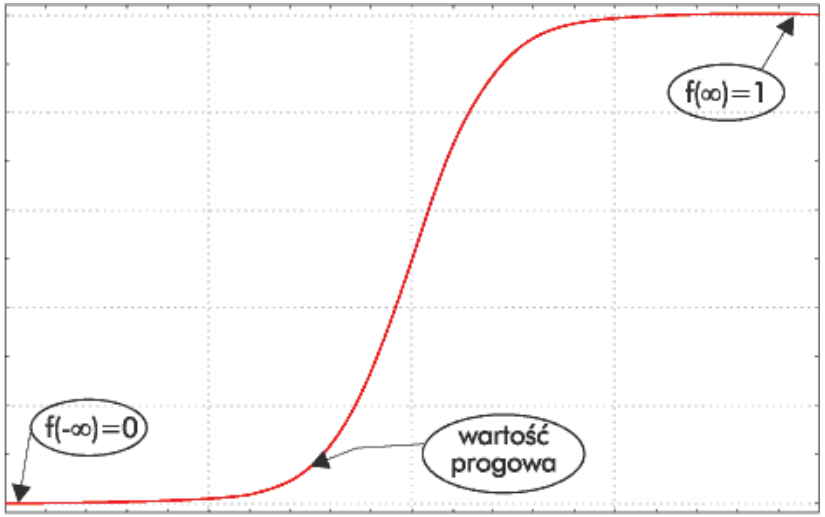
W modelu *regresji logistycznej* (logit), przewidywane wartości zmiennej zależnej nigdy nie będą mniejsze (lub równe) od *0* ani większe (lub równe) od *1*, bez względu na wartości zmiennych niezależnych.

$$\Theta(x) = \frac{\exp(\alpha + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \ldots + \beta_N \cdot x_N)}{1 + \exp(\alpha + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \ldots + \beta_N \cdot x_N)}$$

Parametry równania szacuje się metodą największej wiarygodności, poszukując wartości parametrów α_j maksymalizujących wiarygodność próby, na podstawie której estymuje się model

Miarą wiarygodności jest wyrażenie
-2 lnL (L - funkcja wiarygodności).

Funkcja logistyczna



Funkcja logistyczna przyjmuje wartości od 0 do 1.

Model może opisywać prawdopodobieństwo zachorowania lub szansę wyzdrowienia

- Model wprowadza pewną wartość progową, po przekroczeniu której gwałtownie wzrasta prawdopodobieństwo.
- Model często wykorzystywany w badaniach medycznych
- Szansa

$$(\text{Szansa})\ \mathrm{S(A)} = \frac{p(\mathrm{A})}{p(\mathrm{nieA})} = \frac{p(\mathrm{A})}{1-p(\mathrm{A})}$$

Iloraz szans (poziom szans)

$$\mathrm{OR_{AxB}} = \frac{\mathrm{S(A)}}{\mathrm{S(B)}} = \frac{p(\mathrm{A})}{1-p(\mathrm{A})} : \frac{p(\mathrm{B})}{1-p(\mathrm{B})}$$

Interpretacja

Wartości oszacowanych współczynników nie podlegają interpretacji.

Interpretacji podlega natomiast wyrażenie zwane ilorazem szans:

$$\Psi = \frac{P_i}{1 - P_i} \qquad \Psi = e^{\alpha_0 + \alpha_1 X_1 + \ldots + \alpha_k X_k}$$

Wyrażenie e^{α_j} to relatywna zmiana możliwości wystąpienia zdarzenia pod wpływem czynnika opisanego przez zmienną X_j Stąd, jeśli e^{α_j} <0, to czynnik jest ograniczający, w przeciwnym wypadku – stymulujący

Ocena współczynników regresji

Ocenie podlega także istotność poszczególnych współczynników regresji za pomocą statystyki Walda:

$$\chi^2 = \left(\frac{\hat{\alpha}_j}{S(\hat{\alpha}_j)} \right)$$

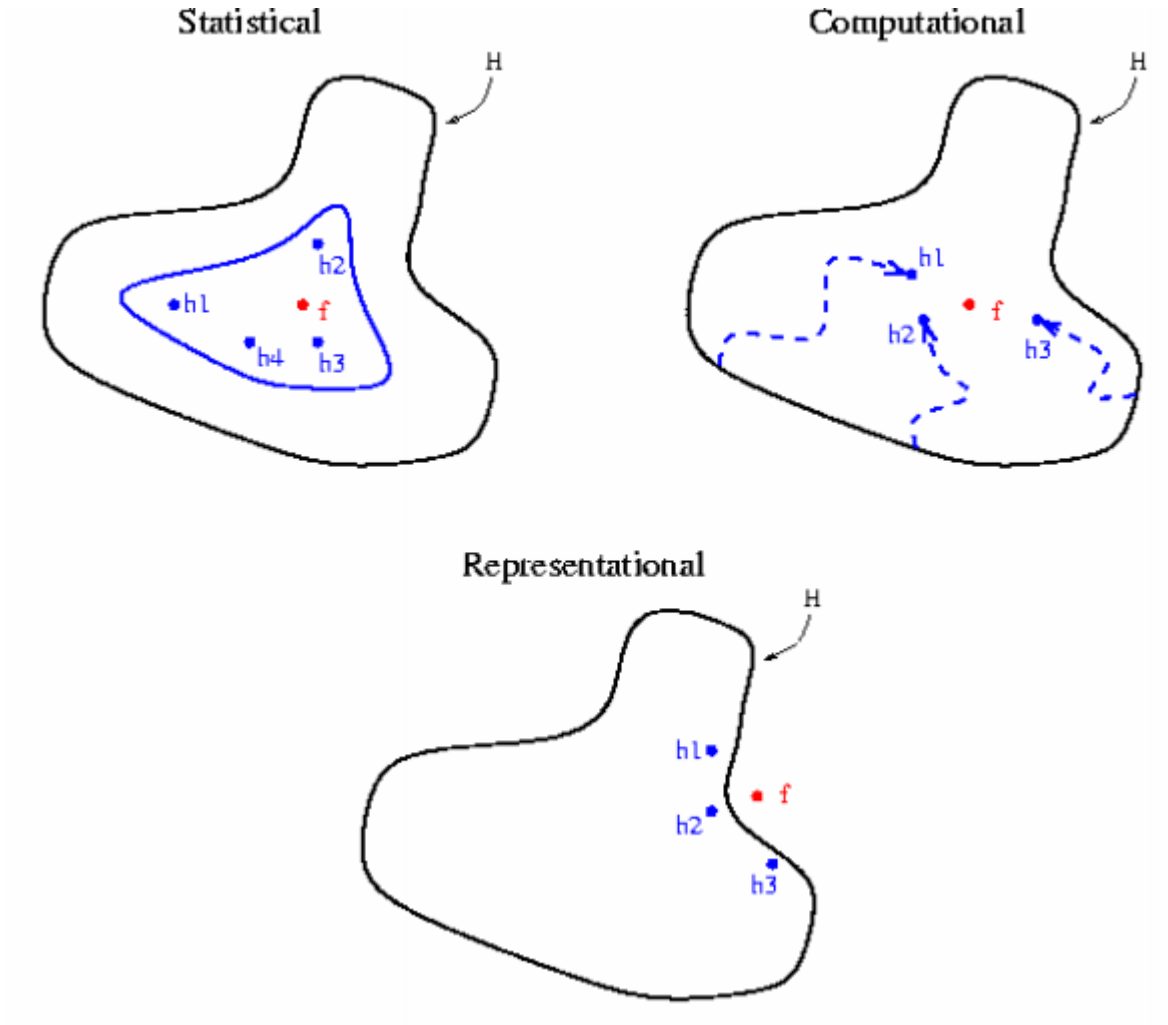
Im wyższa wartość statystyki, tym mocniejsze są podstawy do uznania istotności oszacowanego współczynnika.

multiple classifiers

multistrategy learning

combined approach

Multiple classifiers



Different approaches to create multiple systems

Homogeneous classifiers – use of the same algorithm over diversified data sets

- 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)

Heterogeneous classifiers – different learning algorithms over the same data

- Voting or rule-fixed aggregation
- Stacked generalization or meta-learning: Predictions of base learners (level-0 models) are used as input for meta learner (level-1 model)

some practical advices

prof. Jerzy Stefanowski

If the classifier is unstable (i.e, decision trees) then apply bagging

If the classifier is stable and simple (e.g. Naïve Bayes) then apply boosting

If the classifier is stable and very complex (e.g. Neural Network) then apply randomization injection

If you have many classes and a binary classifier then try errorcorrecting codes.

If it does not work then use a complex binary classifier!

SMOTE

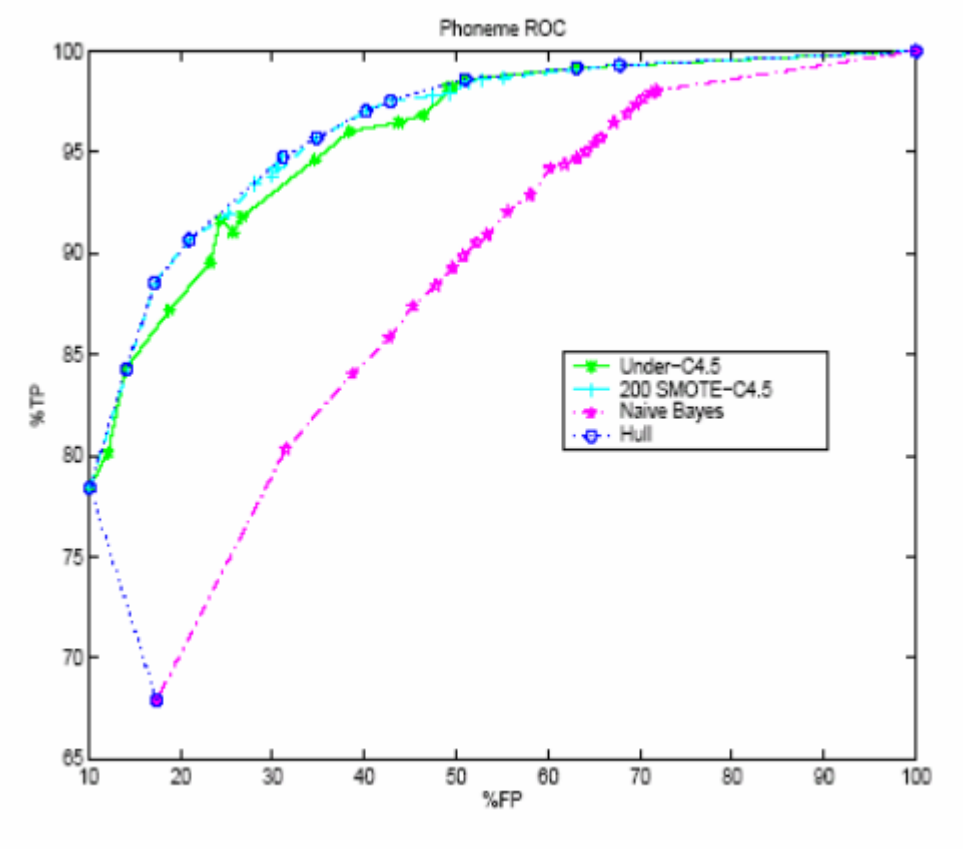
Synthetic Minority Oversampling Technique

- Technique designed by Chawla, Hall, Kegelmeyer 2002
- For each minority Sample
  - Find its k-nearest minority neighbours
  - Randomly select j of these neighbours
  - Randomly generate synthetic samples along the lines joining the minority sample and its j selected neighbours

  (j depends on the amount of oversampling desired)

- Comparing to simple random oversampling – for SMOTE larger and less specific regions are learned, thus, paying attention to minority class samples without causing overfitting.
- SMOTE currently yields the best results as far as re-sampling and combination with undersampling go (Chawla, 2003).

SMOTE - performance



Changing set of rules for the minority class

- Minority class rules have smaller chance to predict classification for new objects!
- Two stage approach (Stefanowski, Wilk):
  1. Induce minimal set of rules for all classes.
  2. Replace the set of rules for the minority class by another set → more numerous and with greater strength.
- The chance of using these rule while classifying new objects is increased.
- The use of EXPLORE (Stefanowski, Vanderpooten):
  - Induce all rules with strength greater then a threshold.
  - Modify the threshold considering gain + conditions calculated from 1 stage.