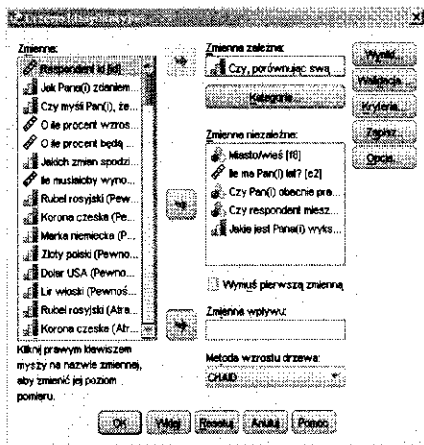
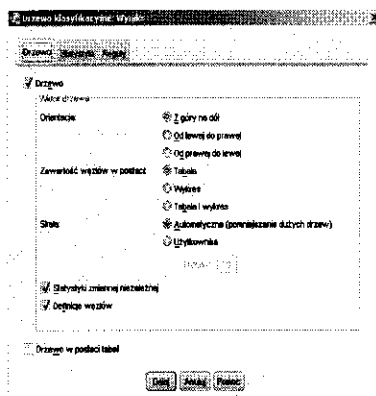


DRZEWA KLASYFIKACYJNE W SPSS – Automatyczna detekcja interakcji chi-kwadrat (CHAID)

Analiza | Klasyfikacja | Drzewo klasyfikacyjne...

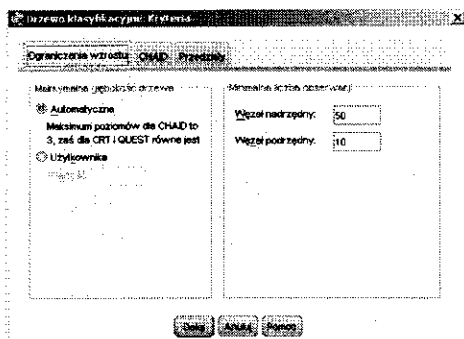


- Wybór zmiennej zależnej (+ możliwość wyboru, które jej kategorie będą analizowane)
- Wybór zmiennej niezależnej (*Wymuś pierwszą zmienną* – pierwszy podział będzie na podstawie pierwszej zmiennej z listy)
- Wybór metody (CHAID, wyczerpujące CHAID, CRT)



Wyniki ...

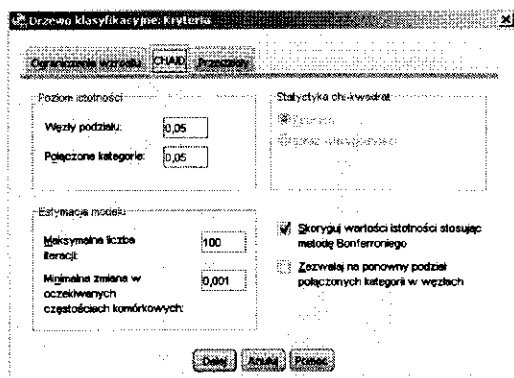
- Sposób prezentacji drzewa
- Czy w węzłach tabela, wykres, czy tabela i wykres
- Czy podawane chi-kwadrat i (skorygowane) p-value
- Czy nad węzłami będzie napisane, czego dotyczy podział



Kryteria ...

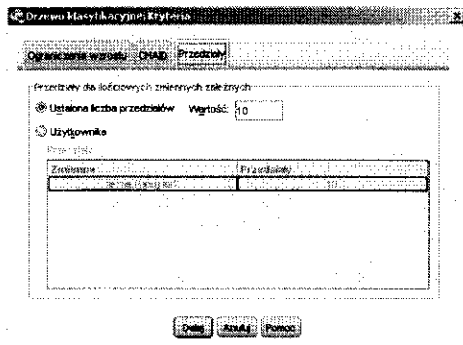
Ograniczenia wzrostu:

- Jaka maksymalna głębokość drzewa (domyślnie dla CHAID 3)
- Ile obserwacji mają co najmniej zawierać węzły (Zmniejszanie powoduje otrzymanie drzewa z większą liczbą węzłów)



CHAID:

- Określenie poziomu istotności (domyślnie 0,05)
- Poprawka Bonferroniego (korekta p-value) - domyślnie stosowana



Przedziały:

- Zmienne metryczne zostają w procedurze przeskalowane na przedziały (domyślnie 10)

Walidacja... pozwala ocenić, jak “dobrze” działa drzewo przy uogólnieniu na większą liczbę obserwacji. Dwie metody walidacji:

- Walidacja krzyżowa
- Walidacja z podziałem próby

Kluczowe elementy raportu:

- a) Model posumowanie – zmienne, algorytm, głębokość, wielkość drzewa
- b) Prezentacja graficzna drzewa

c)

Obserwowana	Macierz klasyfikacji		
	Przewidywana		
	Nie posiada	Posiada	Procent poprawnych
Nie posiada	74	33	69,2%
Posiada	52	169	76,5%
Procent ogółem	38,4%	61,6%	74,1%

Im wyższy procent poprawnie sklasyfikowanych obserwacji, tym lepiej. Można porównać, trafność klasyfikacji dla poszczególnych wariantów zmiennej zależnej

d)

Ryzyko	
Ocena	Błąd standardowy
,259	,024

Ocena ryzyka wskazuje w jakiej części zmienna zależna jest źle przewidywana przez model (w tabeli obok ocena = 0,259, czyli w 25,9% jest źle przewidywana)

Zadanie 1:

- zmienna zależna: czy załatwia(a) Pan(i) kiedykolwiek jakieś sprawy w banku
- zmienne niezależne: wiek, płeć, czy Pan(i) obecnie pracuje, miasto/wieś, jakie jest Pana(i) wykształcenie
 - W wynikach proszę zdefiniować zawartość węzłów w postaci tabeli i wykresu, uwzględnić definicje węzłów i statystyki zmiennej zależnej
 - W kryteriach proszę zdefiniować minimalną liczbę obserwacji dla węzła nadrzędnego 50, dla podrzędnego 10; głębokość drzewa 3; liczbę przedziałów dla zmiennej metrycznej 10.

Proszę określić:

- a) Liczbę liści i je scharakteryzować
- b) Ocenić ryzyko
- c) Ocenić trafność klasyfikacji dla poszczególnych kategorii zmiennej zależnej

Zadanie 2:

- zmienna zależna: credit rating
- zmienne niezależne: wszystkie pozostałe
 - W wynikach proszę zdefiniować zawartość węzłów w postaci tabeli i wykresu, uwzględnić definicje węzłów i statystyki zmiennej zależnej
 - W kryteriach proszę zdefiniować minimalną liczbę obserwacji dla węzła nadrzędnego 400, dla podrzędnego 200; głębokość drzewa 3; liczbę przedziałów dla zmiennej metrycznej 10.
 - Proszę wymedytować drzewo i poddać je obróbce – zmienić kolory czcionki (w węzłach, w definicjach węzłów, zmienić rodzaj czcionki, wypełnić węzły kolorem, zmienić tło, obejrzeć drzewo w poziomie, zmienić tryb wyświetlania – sama tabela, sam wykres, wykres + tabela...)