



## ZASTOSOWANIE TECHNIK ANALIZY SKUPIEŃ I DRZEW DECYZYJNYCH DO SEGMENTACJI RYNKU

*Grzegorz Migut, StatSoft Polska Sp. z o.o.*

Segmentacja rynku jest jednym z kluczowych zadań realizowanych podczas opracowania strategii marketingowych. Dzięki segmentacji możemy wyróżnić grupy klientów podobnych do siebie i co ważniejsze podobnie reagujących na stosowane wobec nich instrumenty marketingowego oddziaływania [1].

Proces identyfikacji i opisu segmentów rynku możemy podzielić na kilka etapów. W pierwszej kolejności konieczne jest określenie kryteriów segmentacji a następnie przeprowadzenie procedury grupowania w oparciu o wybrane kryteria. Kluczowymi elementami tej procedury jest identyfikacja optymalnej liczby segmentów oraz przydzielenie konsumentów do odpowiednich grup. Po wykonaniu grupowania konieczne jest przeprowadzenie opisu poszczególnych segmentów, ocena ich atrakcyjności oraz określenie najbardziej odpowiedniej formy oddziaływania w wybranych segmentach. Oczywiście wyniki przeprowadzonej segmentacji mogą być również znakomitym punktem wyjścia do wykonania pogłębianych analiz dotyczących jedynie wybranych segmentów rynku.

Literatura naukowa i praktyka wyróżniają szereg podziałów i typów segmentacji w zależności od sposobu doboru kryteriów segmentacji i jej celów. My skoncentrujemy się na jednym podstawowym rozróżnieniu na: segmentację opisową oraz segmentację predykcyjną.

W przypadku **segmentacji opisowej** wszystkie zmienne, jakie wykorzystujemy do analizy traktujemy jako zmienne niezależne (wejściowe). Podczas budowy modelu nie mamy żadnego kryterium ukierunkowującego proces poszukiwania jednorodnych grup klientów. Zwykle też na wstępie analizy nie dysponujemy informacją, jakie segmenty występują w danych, ani też ile jest tych segmentów. Wiedzę tę pragniemy dopiero zdobyć w wyniku analizy. Tak sformułowane zadanie analityczne wymaga zastosowania jednej z metod nieukierunkowanego data mining. Najbardziej popularnymi metodami stosowanymi do segmentacji są metody analizy skupień oraz sieci neuronowe Kohonena (SOM).

W przypadku **segmentacji predykcyjnej** wyróżniamy dwa rodzaje zmiennych. Zmienną zależną, która stanowi kryterium segmentacji i jest to najczęściej zmienna opisująca aspekty behawioralne konsumentów [1]. Druga grupa zmiennych to zmienne niezależne, których zadaniem jest wyjaśnienie przejawów zachowania opisanych w kryterium segmentacji. Najczęściej zmiennymi niezależnymi są zmienne demograficzne, geograficzne czy



psychograficzne [1]. Metodą najczęściej wykorzystywaną do przeprowadzenia segmentacji predykcyjnej są drzewa decyzyjne.

## Segmentacja opisowa

Ponieważ podczas segmentacji opisowej nie dysponujemy zmienną, która mogłaby ukierunkować proces identyfikacji segmentów, do analizy wykorzystujemy techniki, które możemy zaliczyć to metod uczenia bez nauczyciela (*unsupervised learning*). Do metod najczęściej wykorzystywanych w segmentacji należą techniki analizy skupień oraz sieci neuronowe Kohonena.

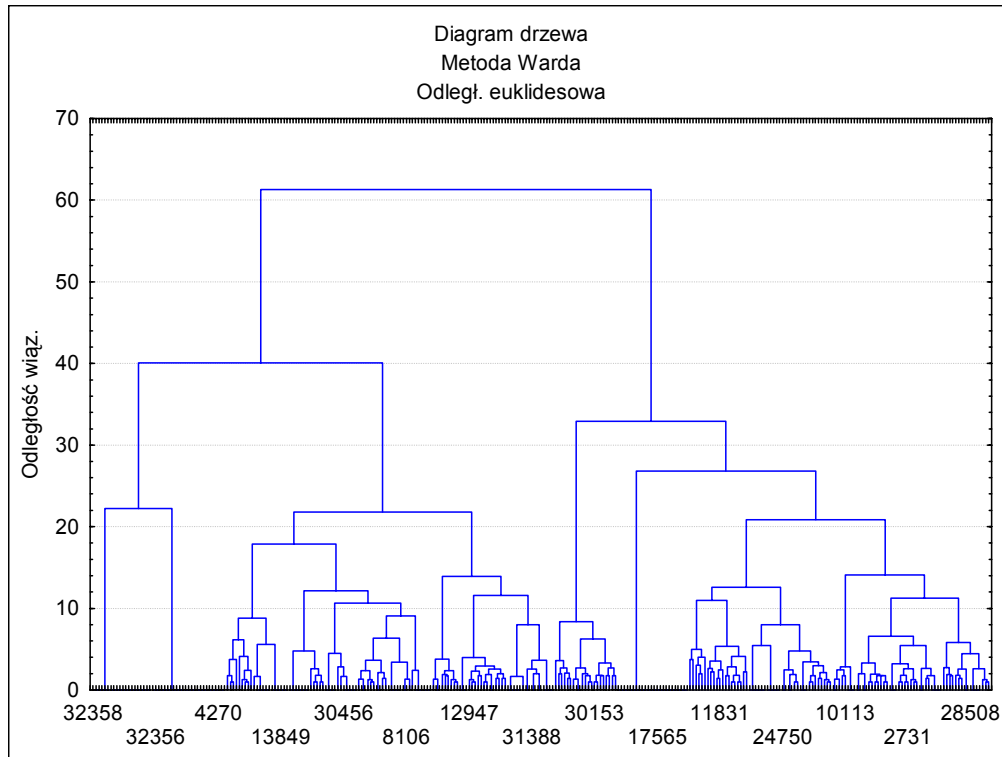
### *Analiza skupień*

Celem **analizy skupień** (*cluster analysis*) jest wyodrębnienie ze zbioru danych obiektów, które byłyby podobne do siebie, i łączenie ich w grupy. W wyniku działania tej analizy z jednego niejednorodnego zbioru danych otrzymujemy grupę kilku jednorodnych podzbiorów. Obiekty znajdujące się w tym samym zbiorze uznawane są za „podobne do siebie”, obiekty z różnych zbiorów traktowane są jako „niepodobne”. Techniki analizy skupień obejmują kilka różnych algorytmów, które można najogólniej podzielić na metody hierarchiczne i niehierarchiczne.

Hierarchiczną metodą analizy skupień jest metoda aglomeracyjna. Algorytm aglomeracji służy do grupowania obiektów w coraz to większe zbiory (skupienia), z zastosowaniem pewnej miary podobieństwa lub odległości. Typowym wynikiem tego typu grupowania jest hierarchiczne drzewo (zob. rysunek). Na początku tej analizy uznajemy, że każdy element zbioru stanowi oddzielną grupę. Następnie stopniowo osłabiamy kryterium uznawania obiektów za takie same, co powoduje grupowanie się obiektów podobnych. W miarę dalszego osłabiania kryterium wiążemy ze sobą coraz więcej obiektów i agregujemy je w coraz większe skupienia elementów, coraz bardziej różniących się od siebie. W końcu, na ostatnim etapie, wszystkie obiekty zostają ze sobą połączone. Efekty działania tego algorytmu można przedstawić w formie hierarchicznego drzewa, które przedstawia kolejne kroki działania analizy.

Tego typu analizę możemy przeprowadzić nie tylko dla przypadków, ale również dla zmiennych, co polega na łączeniu najbardziej podobnych zmiennych (w sensie odległości, a nie korelacji) w grupy, podobnie jak przedstawiono poniżej. Ważnym parametrem wpływającym na jakość procesu grupowania jest wybór metody aglomeracji, czyli sposobu liczenia odległości pomiędzy skupieniami (według wybranej metryki).

Odległość między skupieniami może być liczona jako odległość między najbliższymi (pojedyncze wiązanie) lub najdalszymi (pełne wiązanie) reprezentantami poszczególnych skupisk, bądź też na podstawie średnich, median lub środków ciężkości skupień (ważonych lub nieważonych). Metoda Warda (minimum wariancji) wykorzystuje w wyodrębnianiu skupisk zasadę minimalizacji wariancji wewnątrzklasowej.

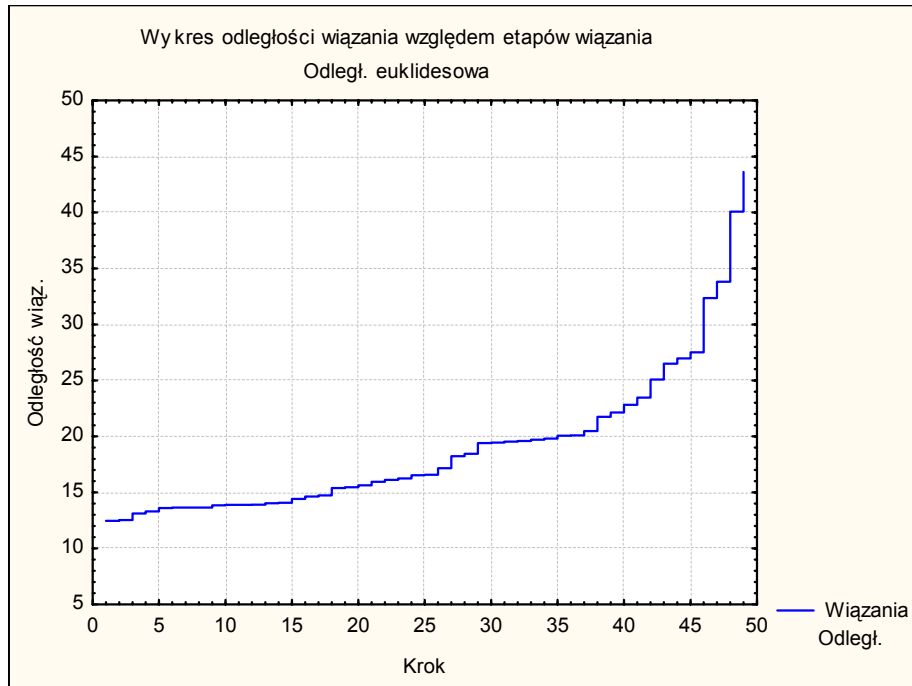


W tej ostatniej metodzie wyodrębniania skupisk powinna być stosowana odległość euklidesowa lub kwadratowa odległość euklidesowa [1]. Spośród wymienionych metod najbardziej godna polecenia ze względu na kryterium efektywności odtwarzania rzeczywistej struktury danych jest metoda Warda [2].

Metoda aglomeracyjna jest rzadko stosowana w segmentacji dużej liczby obiektów (powyżej 300), ponieważ wymaga obliczenia macierzy odległości pomiędzy wszystkimi analizowanymi obiektami, co jest bardzo wymagające numerycznie. Jeszcze bardziej ograniczającym kryterium jest kwestia czytelności wykresu aglomeracji, który traci przejrzystość przy większej liczbie obiektów. Metoda ta jest jednak bardzo pomocna podczas ustalania optymalnej liczby skupień, na jaką należy podzielić analizowaną zbiorowość.

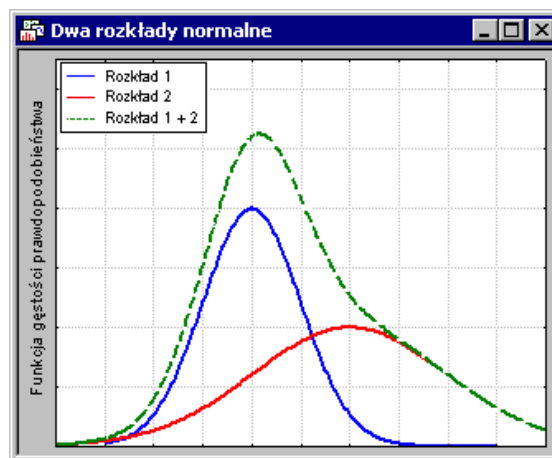
Określając optymalną liczbę segmentów na podstawie analizy aglomeracyjnej, możemy posłużyć się na przykład kryterium maksymalnego ilorazu odległości aglomeracyjnych wykorzystywanych w dwóch sąsiednich aglomeracjach. Innym kryterium może być pierwszy wyraźny przyrost odległości aglomeracyjnej [2], który możemy zaobserwować, analizując wykres odległości aglomeracyjnej dla kolejnych etapów wiązania (zob. rysunek poniżej).

Do najważniejszych metod niehierarchicznych należy zaliczyć metodę  $k$ -średnich oraz EM. Stosowanie metody  $k$ -średnich wymaga podania liczby grup, na które zostanie podzielony wejściowy zbiór danych. Jedną z wersji tej metody polega na losowym wyborze  $k$ -obiektów z analizowanego zbioru i uznania ich za środki  $k$  grup. Każdy z pozostałych obiektów jest przypisywany do grupy o najbliższym mu środku. Następnie oblicza się nowe środki każdej podgrupy na podstawie średnich arytmetycznych ze współrzędnych zawartych w nich obiektów.



W kolejnym kroku następuje przegrupowanie elementów grup, każdy obiekt jest przesuwany do tej grupy, do której środka ma najbliższej. Procedurę tę powtarzamy do momentu, gdy w danej iteracji żaden z obiektów nie zmieni swojej podgrupy. Pewną wadą tej metody jest konieczność odgórnego określenia liczby skupień występujących w danych, dlatego też zaleca się powtórzenie procedury dla różnych wartości  $k$  i wybranie tej, dla której zbiór danych jest podzielony najlepiej, lub oparcie się na wynikach analizy aglomeracyjnej.

Metoda EM jest czasem nazywana analizą skupień bazującą na prawdopodobieństwie lub statystyczną analizą skupień. Program wyznacza skupienia, zakładając różne rozkłady prawdopodobieństwa zmiennych uwzględnianych w analizie. Na początku działania algorytmu, podobnie jak w metodzie  $k$ -średnich, musimy podać liczbę skupień, jakie powinny być wyodrębnione ze zbioru wejściowego.



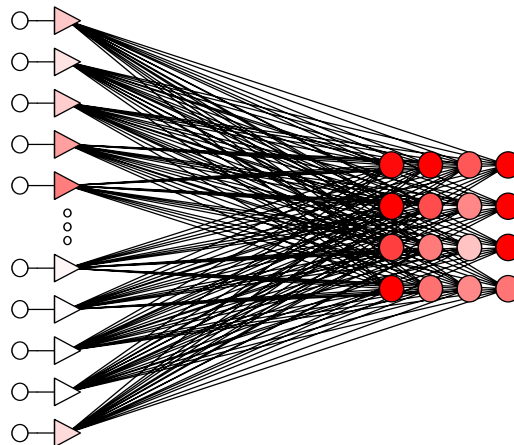
Założmy, że przeprowadziliśmy badania w pewnej dużej zbiorowości pod kątem jednej cechy ciągłej. Zaobserwowany rozkład tej cechy był zgodny z funkcją gęstości opisaną



linią przerywaną (*Rozkład 1+2*) charakteryzującą się pewną średnią oraz odchyleniem standardowym. Wiemy też, że w zbiorowości tej występują dwa segmenty (na przykład kobiety i mężczyźni) o różnych parametrach funkcji gęstości w swoich segmentach. Algorytm EM ma na celu określenie parametrów rozkładów segmentów na podstawie rozkładu całej grupy oraz przydzielenie poszczególnych obserwacji do najbardziej odpowiadających im segmentów (klasyfikacja następuje na zasadzie prawdopodobieństwa). Na naszym rysunku rozkłady dwóch segmentów zostały przedstawione jako *Rozkład 1* oraz *Rozkład 2*. Po zsumowaniu dają one funkcję rozkładu całej zbiorowości (*Rozkład 1+2*). Algorytm EM przeprowadza klasyfikację nie tylko przy założeniu normalności rozkładu, jak to zaprezentowano na rysunku. Wykorzystując go, można również określić inną funkcję gęstości dla badanej cechy (badanych cech).

### **Sieci neuronowe - sieć Kohonena (*Self Organizing Map*)**

**Sieć Kohonena (SOM)** została zaprojektowana do uczenia w trybie bez nauczyciela – podczas uczenia ustalanie parametrów sieci nie jest sterowane za pomocą wartości wyjściowych, podczas nauki prezentowane są jedynie dane kierowane na wejścia sieci. Sieć ta posiada dwie warstwy: warstwę wejściową oraz warstwę wyjściową składającą się z neuronów radialnych. Warstwa ta znana jest również jako warstwa tworząca mapę topologiczną, ponieważ takie jest jej najczęstsze zastosowanie. Neurony w warstwie tworzącej mapę topologiczną zwykle wyobrażamy sobie jako węzły dwuwymiarowej siatki, chociaż możliwe jest również tworzenie jednowymiarowych sieci w postaci długich łańcuchów.



Sieci Kohonena uczone są przy wykorzystaniu algorytmu iteracyjnego. Rozpoczynając od początkowych, wybranych w sposób losowy centrów radialnych, algorytm stopniowo modyfikuje je w taki sposób, aby odzwierciedlić skupienia występujące w danych uczących. Iteracyjna procedura ucząca dodatkowo porządkuje neurony reprezentujące centra położone blisko siebie na mapie topologicznej.



Podstawowy iteracyjny algorytm działa przez dużą liczbę epok (podczas każdej epoki prezentowany jest sieci cały zestaw danych) w następujący sposób [3]:

- ◆ Pokazywany jest zestaw danych wejściowych ze zbioru uczącego.
- ◆ Wszystkie neurony sieci wyznaczają swoje sygnały wyjściowe, stanowiące odpowiedź na podane wejścia.
- ◆ Wybierany jest neuron zwycięski (tzn. ten, który reprezentuje centrum najbardziej zbliżone do prezentowanego na wejściu przypadku).
- ◆ Neuron zwycięski modyfikowany jest w taki sposób, aby upodobnić jego wzorec do prezentowanego przypadku. W tym celu wyznaczana jest ważona suma przechowywanego w neuronie centrum oraz przypadku uczącego.
- ◆ Wraz ze zwycięskim neuronem w podobny sposób modyfikowane są parametry jego sąsiadów (sąsiedzi wyznaczani są w oparciu o przyjęty wzór topologii sieci).

Algorytm wykorzystuje zmienny w czasie współczynnik uczenia, który jest używany do wyznaczenia ważonej sumy, i powoduje, że zmiany - początkowo duże i szybkie - stają się coraz bardziej subtelne w trakcie kolejnych epok. Umożliwia to ustalenie centrów w taki sposób, że stanowią one pewien kompromis pomiędzy wieloma przypadkami powodującymi zwycięstwo rozważanego neuronu.

Własność uporządkowania topologicznego jest osiągnięta przez zastosowanie w algorytmie koncepcji sąsiedztwa. Sąsiedztwo stanowią neurony otaczające neuron zwycięski. Sąsiedztwo, podobnie jak współczynnik uczenia, zmniejszane jest wraz z upływem czasu, tak więc początkowo do sąsiedztwa należy stosunkowo duża liczba neuronów; w końcowych etapach sąsiedztwo ma zerowy zasięg. Ma to istotne znaczenie, ponieważ w algorytmie Kohonena modyfikacja wag jest w rzeczywistości przeprowadzana nie tylko w odniesieniu do neuronu zwycięskiego, ale również we wszystkich neuronach należących do sąsiedztwa.

Po nauczeniu sieci Kohonena poprawnego rozpoznawania struktury prezentowanych danych można jej użyć jako narzędzia przeprowadzającego wizualizację danych w celu ich lepszego poznania. Ważnym elementem przygotowania sieci Kohonena do bieżącego użytkowania jest właściwe opisanie uformowanej mapy topologicznej. Ustalenie związków pomiędzy skupieniami a znaczeniami wymaga zwykle odwołania się do dziedziny, której dotyczy analiza [3].

### ***Przykład wykorzystania metody k-średnich do segmentacji opisowej***

Do budowy przykładowego modelu segmentacyjnego użyjemy pliku *Adults.sta*, który jest nieco zmodyfikowaną wersją danych dostępnych w Internecie<sup>9</sup>.

Analizowane dane zawierają ponad 32 tysiące przypadków, każdy z nich reprezentuje jedną osobę. Każda z osób opisana jest przez 11 cech demograficznych, takich jak: wiek, płeć, stan cywilny czy grupa zawodowa. Naszym celem będzie wyróżnienie optymalnej z punktu

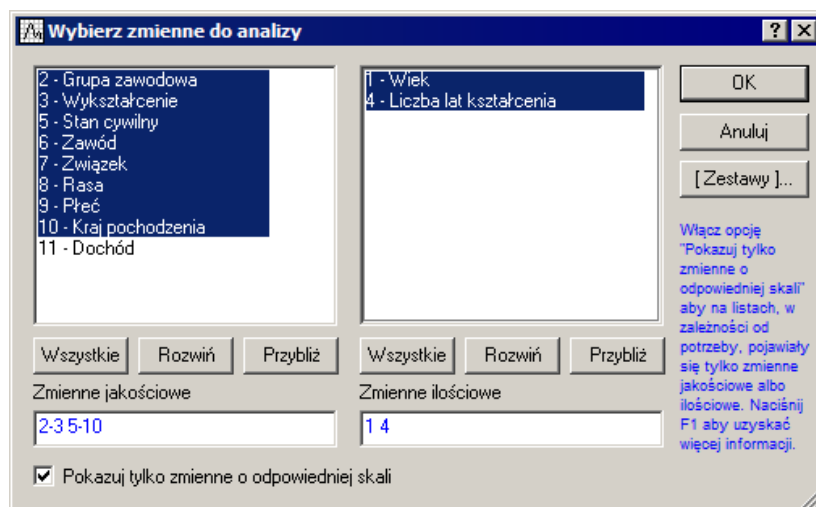
<sup>9</sup> Na przykład *UCI Machine Learning Repository* - <http://archive.ics.uci.edu/ml/>.



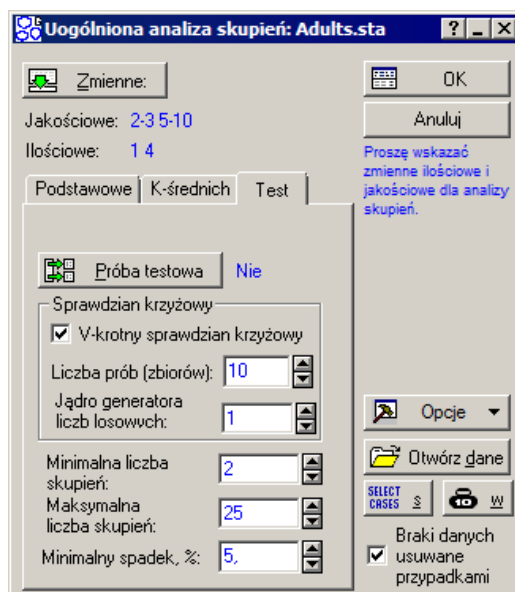
widzenia analizy liczby segmentów, dokonanie klasyfikacji osób do poszczególnych grup oraz ich późniejsza charakterystyka.

Do przeprowadzenia segmentacji użyjemy uogólnionej metody *k*-średnich zaimplementowanej w programie *STATISTICA Data Miner*. Metoda ta, oprócz wbudowanej obsługi cech jakościowych (klasyczne metody analizy skupień operują jedynie na zmiennych ilościowych), pozwala również na automatyczną identyfikację optymalnej liczby segmentów. Mechanizm ten dostępny jest dla metody *k*-średnich oraz EM i jest oparty na *v*-krotnym teście krzyżowym. Algorytm działa w ten sposób, że dzieli zbiór wejściowy na kolejno coraz większą liczbę segmentów i ocenia precyzję podziału dla każdego z nich. Jeśli w wyniku kolejnego podziału zbudowany model poprawia się w stosunku do poprzedniego modelu w mniejszym stopniu niż określono to w wartości progowej (domyślnie jest to 5%), algorytm zatrzymuje swoje działanie (dodanie kolejnego segmentu w znaczący sposób nie poprawia wyników modelu). Dla metody *k*-średnich miarą precyzji podziału jest przeciętna odległość elementów zbioru wejściowego od środka segmentu, w jakim się znajdują, w przypadku metody EM miara ta bazuje na prawdopodobieństwie przynależności do odpowiednich segmentów.

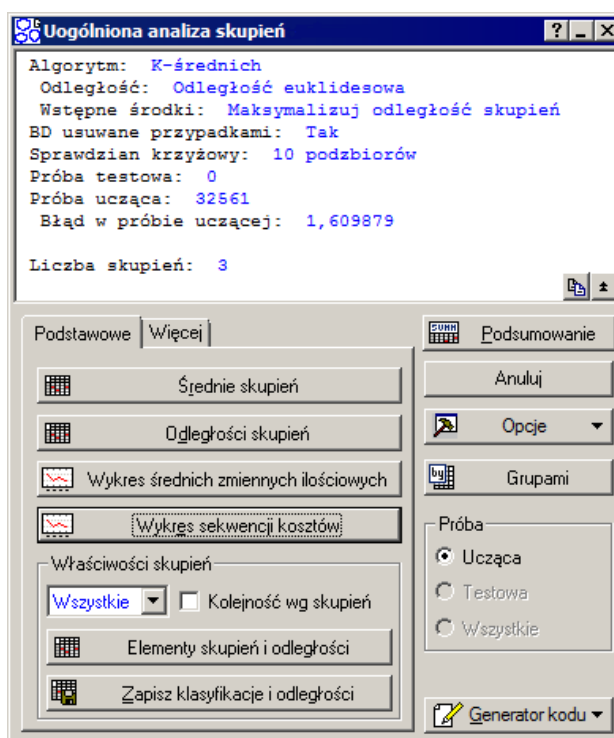
Aby rozpocząć proces analizy, z menu *Data Mining* wybieramy opcję *Analiza skupień uogólnioną metodą EM i k-średnich*, a następnie w wyświetlonym oknie klikamy przycisk *Zmienne*, aby wybrać zmienne do analizy.



Po zaznaczeniu opcji *Pokazuj tylko zmienne o odpowiedniej skali* na liście *Zmienne jakościowe* wybieramy wszystkie zmienne, poza zmienną *Dochód* (użyjemy jej dalszej części analizy). Na liście *Zmienne ilościowe* wybieramy zmienne *Wiek* oraz *Liczba lat kształcenia*. Następnie przechodzimy na kartę *Test* i zaznaczamy opcję *V-krotny sprawdzian krzyżowy*.



Opcja ta pozwoli automatycznie określić optymalną (z punktu widzenia danych) liczbę segmentów. Po jej zaznaczeniu zatwierdzamy wykonanie analizy, klikając **OK**.

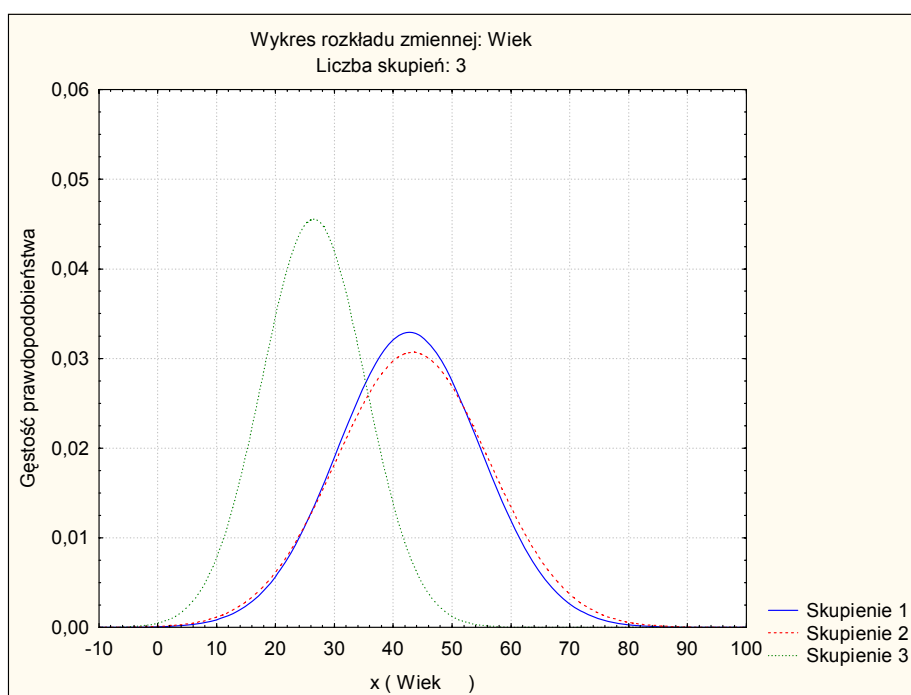


Po wykonaniu analizy możemy stwierdzić, że algorytm wyróżnił trzy skupienia. Kluczowym zadaniem na tym etapie analizy jest odpowiednie scharakteryzowanie uzyskanych skupień. W pierwszej kolejności klikamy przycisk *Średnie skupień*, by wygenerować raport zawierający informację o licznosciach poszczególnych segmentów oraz średnich wartościach analizowanych cech ilościowych.



Skupienie				
	Wiek	Liczba lat kształcenia	Liczba przypadków	Procent (%)
1	42,7	10,6	16362	50,3
2	43,3	9,4	7623	23,4
3	26,5	9,7	8576	26,3

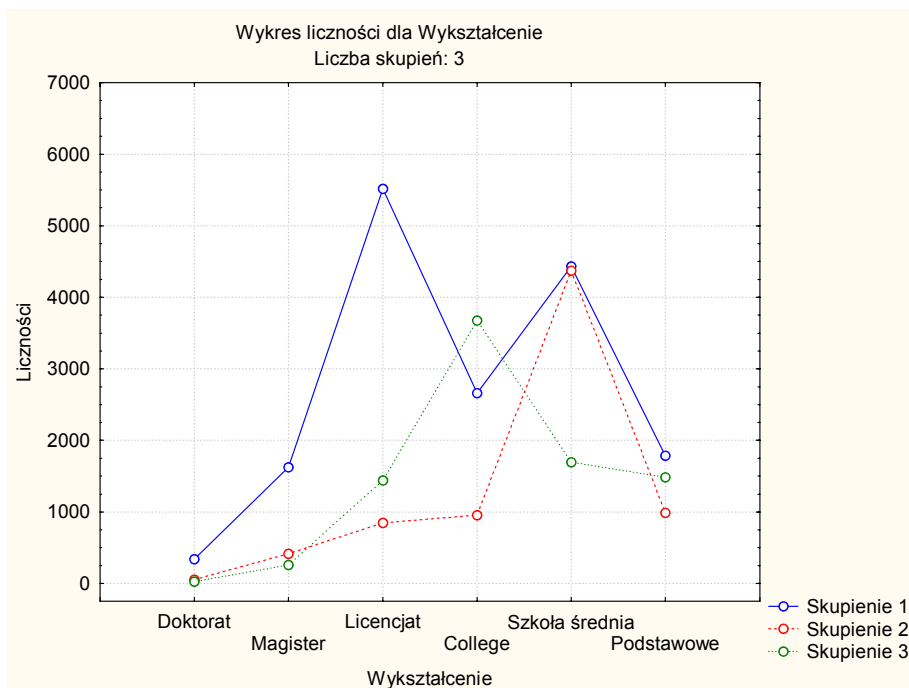
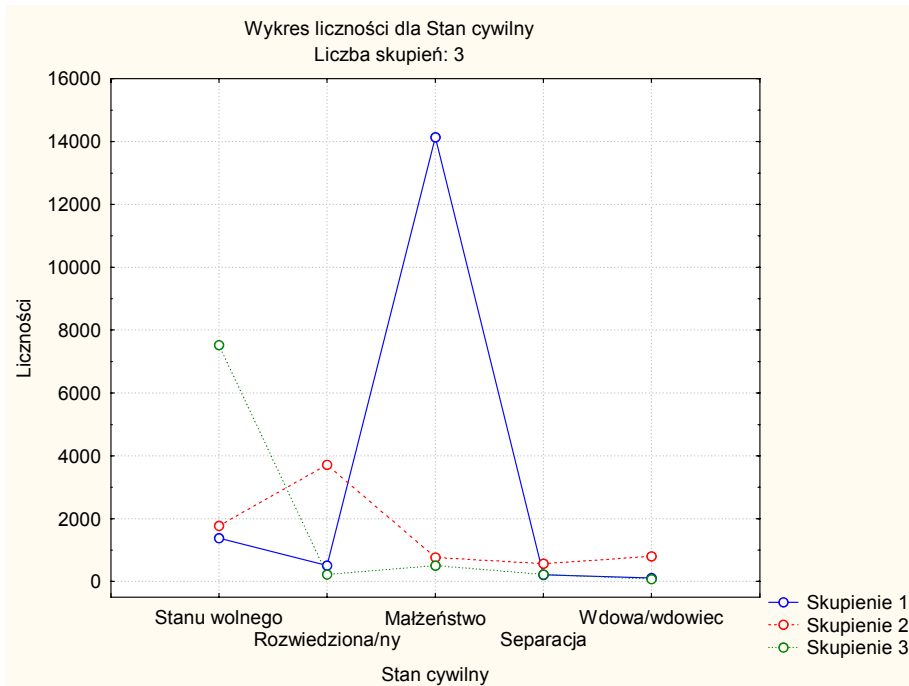
Analizując powyższy raport możemy zauważyć, iż do segmentu pierwszego trafiła nieco ponad połowa analizowanych osób. Segment 2 to nieco ponad 23%, natomiast segment trzeci reprezentuje 26% osób. Analizując średnie wartości zmiennych ilościowych, możemy zauważyć, że osoby z segmentu trzeciego mają średnio 26,5 roku, znacznie mniej niż osoby z pierwszego i drugiego segmentu. Cecha *Liczba lat kształcenia* nie różnicuje zasadniczo osób w poszczególnych segmentach.



Różnice w wieku osób, które trafiły do poszczególnych segmentów, dobrze oddaje również powyższy wykres, na którym możemy ocenić rozkład zmiennej *Wiek* w zależności od określonego segmentu.

Aby scharakteryzować segmenty pod kątem cech jakościowych, przechodzimy na kartę *Więcej* i klikamy przycisk *Wykresy licznosci*, co spowoduje wygenerowanie wykresów opisujących profil segmentów w kontekście poszczególnych zmiennych.

Poniżej zamieszczono dwa przykładowe wykresy licznosci dla cech *Stan cywilny* oraz *Wykształcenie*.



Analizując powyższe wykresy, możemy stwierdzić, że w pierwszym segmencie dominują osoby żyjące w małżeństwie mające najwyższe wykształcenie przy czym dominującym jest *Licencjat*. Osoby znajdujące się w segmencie drugim to bardzo często osoby rozwiedzione, częściej niż w przypadku innych segmentów trafiają też do niego wdowcy. Osoby te najczęściej kończyły edukację na poziomie szkoły średniej.

Segment trzeci to najczęściej osoby stanu wolnego, dominujący poziom wykształcenia to *College*, choć sądząc po średnim wieku tego segmentu, może on w przyszłości wykazać znaczną dynamikę zmian w tym względzie. Oczywiście by jeszcze pełniej scharakteryzować



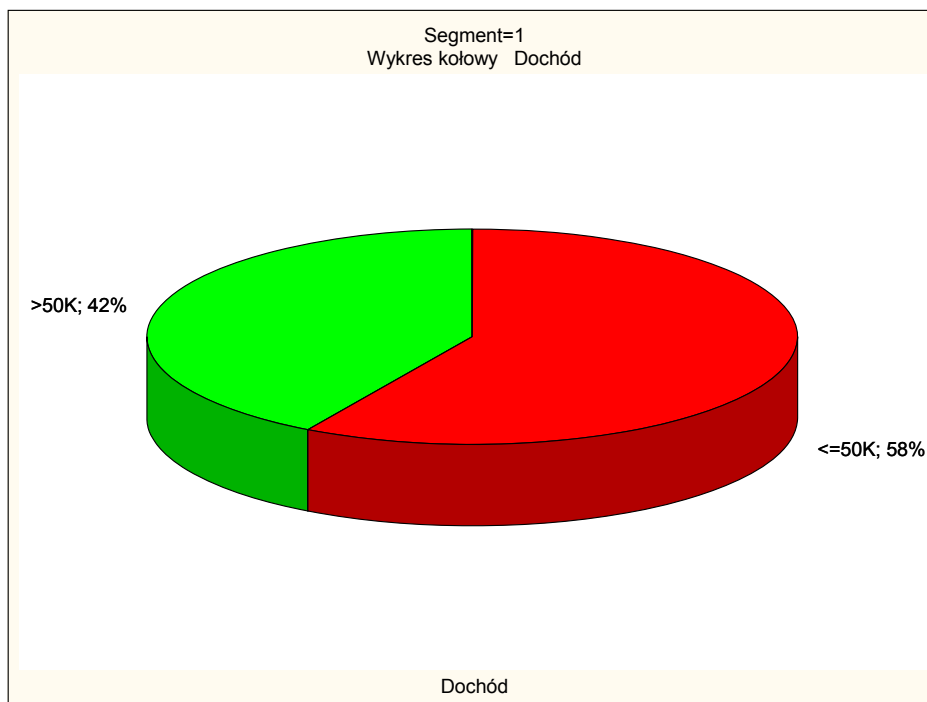
uzyskane segmenty, warto byłoby ocenić profile pozostałych zmiennych jakościowych. My jednak użyjemy do uzupełnienia opisu segmentów dodatkowej zmiennej profilującej *Dochód*, której nie uwzględniliśmy w procesie identyfikacji segmentów.

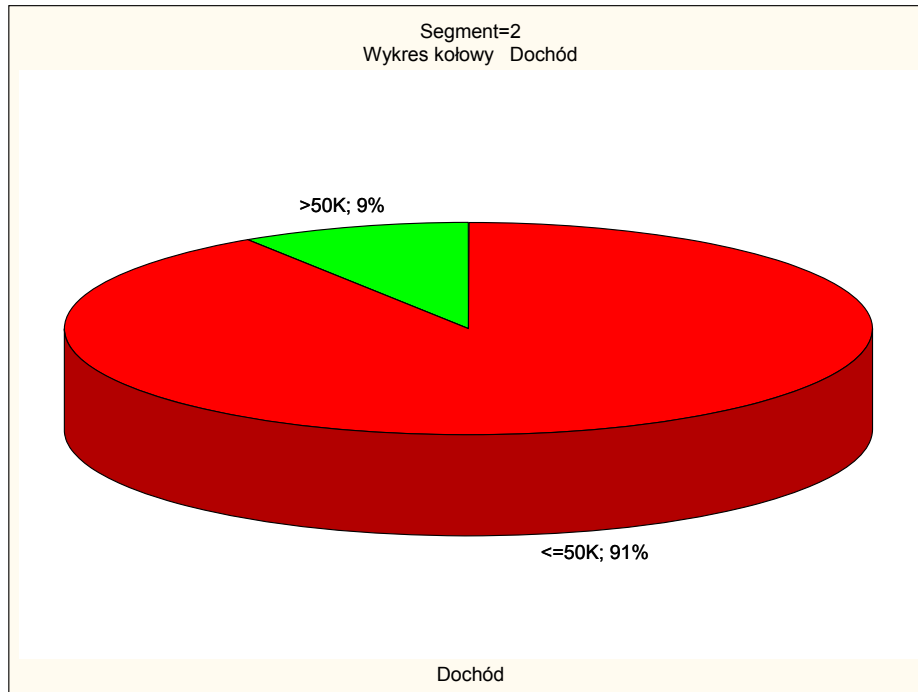
Aby móc to wykonać, za pomocą przycisku *Zapisz klasyfikacje i odległości* utworzymy arkusz danych, który obok zmiennych zawartych w wejściowym arkuszu danych zawierał będzie zmienną informującą o numerze segmentu, do którego została przypisana dana osoba. Dzięki tej operacji zmienną *Dochód*, która nie została użyta podczas analizy, możemy wykorzystać jako dodatkową zmienną opisującą nasze segmenty. Zmienna *Dochód* jest zmienną jakościową przyjmującą dwie wartości  $\leq 50K$ , jeżeli dochód danej osoby nie przekracza 50 tys. \$, oraz  $> 50K$ , jeżeli dana osoba zarabia powyżej tej kwoty.

Do oceny segmentów pod kątem zmiennej *Dochód* użyjemy wykresów kołowych. Z menu *Wykresy* wybieramy opcję *Wykresy 2W -> Wykresy kołowe*, a następnie wskazujemy zmienną *Dochód* jako zmienną do analizy. Dodatkowo za pomocą opcji *Grupami* wskażemy zmienną *Wynikowa klasyfikacja* (zawierającą informację o numerze segmentu), co pozwoli nam ocenić różnice w rozkładzie zmiennej *Dochód* w poszczególnych segmentach.

Analizując poniższe wykresy, możemy zauważyć, że w segmencie 1 znaczny odsetek osób (42%) to osoby o dochodzie powyżej 50 tys. \$. W segmencie 2 osoby te stanowią jedynie 9%, natomiast w segmencie 3 (nie zamieszczonym na rysunku) jedynie 4% osób.

Dodatkowo, jeśli chcielibyśmy przypisać do wyróżnionych segmentów nowe przypadki, możemy to zrobić po wygenerowaniu modelu segmentacji w postaci pliku PMML za pomocą przycisku *Generator kodu*. Model ten możemy następnie uruchomić dla nowych danych za pomocą modułu *Szybkie wdrażanie modeli predykcyjnych*.





## Segmentacja predykcyjna

Segmentacja predykcyjna oparta jest na kryteriach segmentacyjnych, które są wyjaśniane za pomocą dodatkowych zmiennych. Kryteria te charakteryzują najczęściej aspekty behawioralne konsumentów. Zachowanie klientów zostaje wyjaśnione za pomocą odpowiednich zmiennych niezależnych, do których należą zmienne demograficzne, geograficzne czy psychograficzne [1].

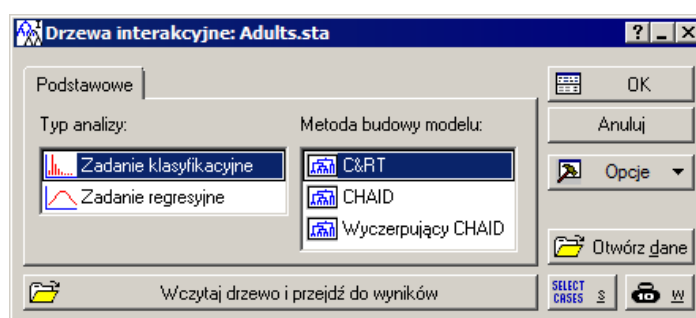
Metodą najczęściej wykorzystywaną do przeprowadzenia segmentacji predykcyjnej są drzewa decyzyjne. Proces budowy drzewa opiera się na zasadzie rekurencyjnego podziału. Zasada ta polega na przeszukiwaniu w przestrzeni cech wszystkich możliwych podziałów zbioru danych na dwie części, tak by dwa otrzymane podzbiory maksymalnie się między sobą różniły ze względu na zmienną zależną (kryterium segmentacji). Podział ten jest kontynuowany, aż do podziału przypadków na jednorodne grupy lub spełnienia ustalonych warunków zatrzymania. Reguły, względem których dokonano podziału przestrzeni cech, można w łatwy sposób przedstawić w formie drzewa. Tego typu grafy składają się z wierzchołków i krawędzi. Każdy wierzchołek reprezentuje decyzję o podziale zbioru obiektów na dwa podzbiory ze względu na jedną z cech objaśniających. Z drzewami klasyfikacyjnymi związane jest też pojęcie przycinania. Jest ono użyteczne do zoptymalizowania struktury drzewa. Przycinanie polega na upraszczaniu struktury drzewa (usuwanie węzłów) przy równoczesnym zachowaniu (ewentualnie dopuszczalnym przez nas pogorszeniu) jego zdolności klasyfikacyjnych. Ważną zaletą drzew jest zrozumiała dla człowieka sekwencja reguł decyzyjnych pozwalająca klasyfikować nowe obiekty na podstawie wartości zmiennych. Atrakcyjną jest również możliwość graficznej prezentacji procesu klasyfikacji. Dodatkową zaletą drzew klasyfikacyjnych jest ich odporność na obserwacje odstające.



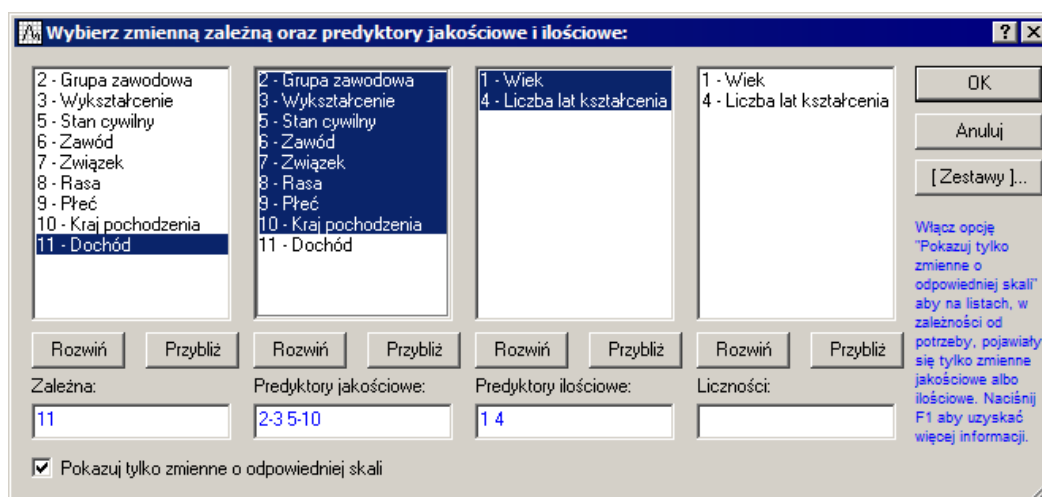
## Przykład wykorzystania metody drzew CART do segmentacji predykcyjnej

Do segmentacji predykcyjnej użyjemy tego samego zbioru danych, na podstawie którego przygotowaliśmy segmentację opisową, z tą różnicą, że tym razem wykorzystamy dodatkowe kryterium segmentacji, jakim będzie fakt uzyskiwania przychodów na poziomie wyższym bądź niższym od 50 tys. dolarów. Fakt ten zapisany został w zmiennej *Dochód*, która w naszej analizie pełniła będzie rolę zmiennej zależnej. Pozostałe zmienne, które analizowaliśmy poprzednio, będziemy traktować jako zmienne niezależne.

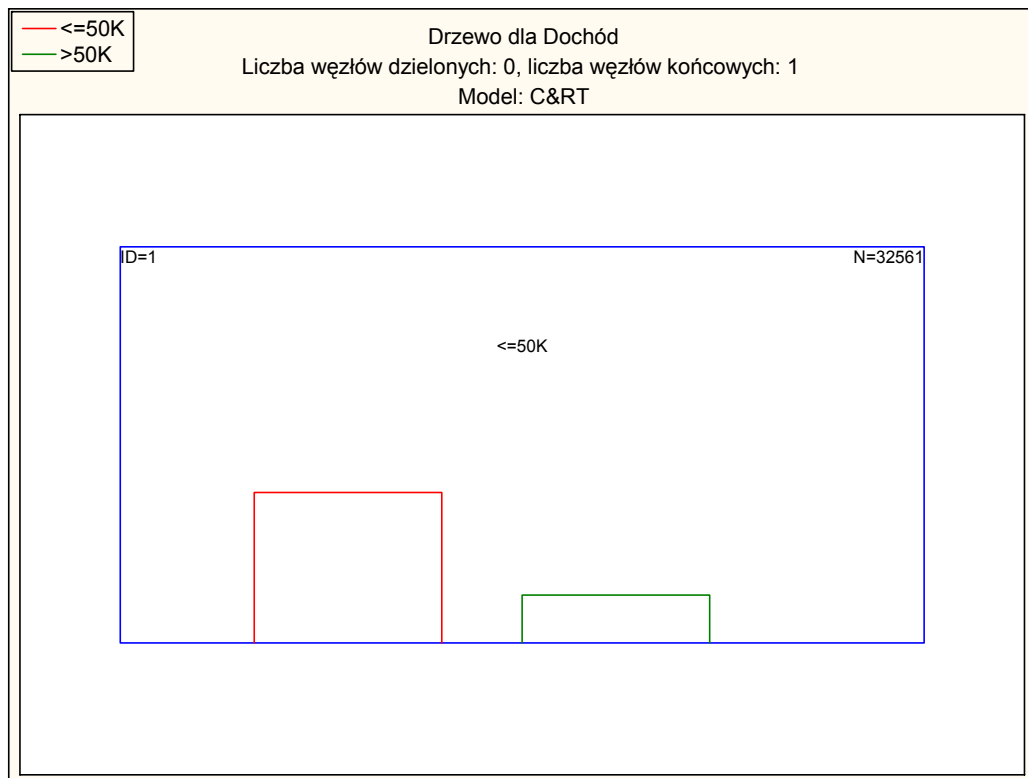
Aby rozpocząć analizę z menu *Data Mining* wybieramy polecenie *Drzewa interakcyjne (C&RT, CHAID)*, a następnie określamy typ zadania jako *Zadanie klasyfikacyjne* (zmienna *Dochód* jest zmienną jakościową), natomiast metodą, jakiej użyjemy do analizy, będą drzewa CART (*C&RT*).



Po zatwierdzeniu wstępnych ustawień analizy w kolejnym kroku musimy określić zmienne, jakie będziemy analizować. Klikamy przycisk *Zmienne* znajdujący się na karcie *Podstawowe* i w oknie wyboru zmiennych wskazujemy zmienną *Dochód* jako zmienną zależną, zmienne *Wiek* oraz *Liczba lat kształcenia* jako predyktory ilościowe, pozostałe zmienne określamy jako predyktory jakościowe.



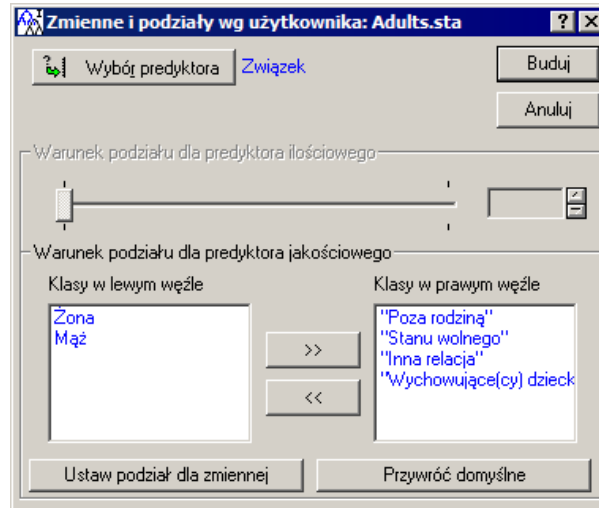
Pozostawiamy domyślne ustawienia pozostałych parametrów analizy i przechodzimy do okna *Wyniki*. W tym momencie rozpoczynamy interakcyjne budowanie modelu oraz identyfikowanie czynników istotnie wpływających na poziom dochodu.



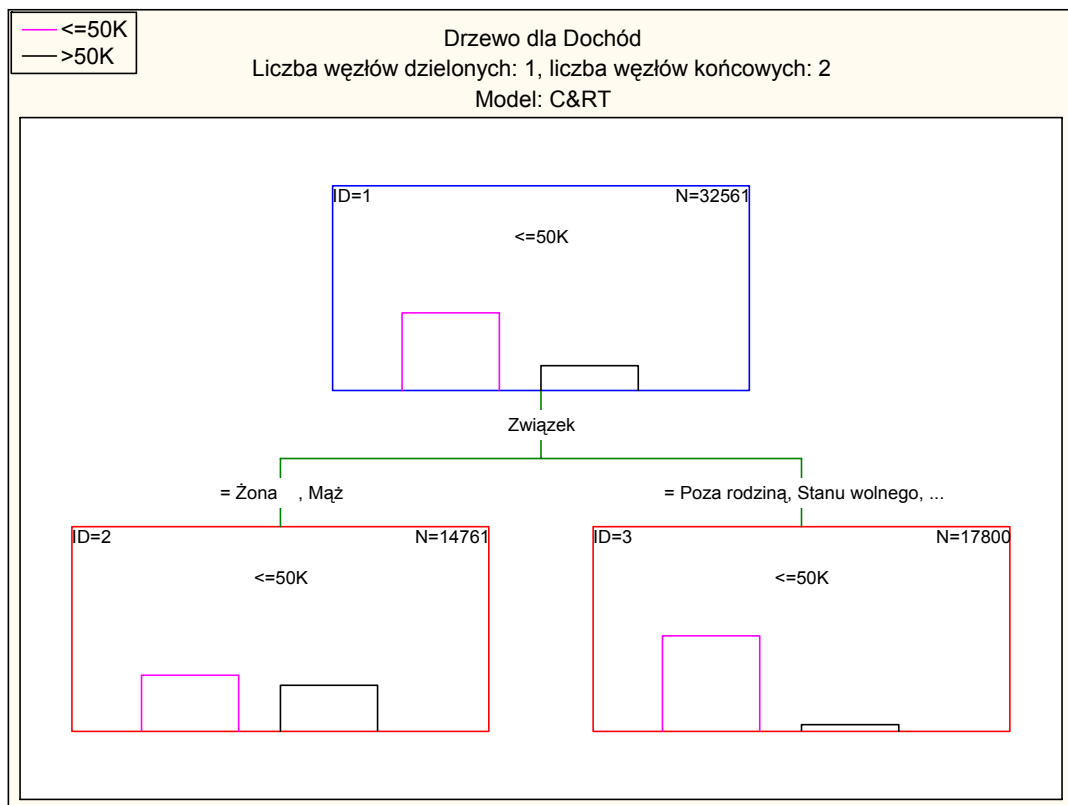
Na początku analizy nasze drzewo składa się jedynie z węzła macierzystego (rysunek powyżej), który nie ma jeszcze żadnych węzłów potomnych. Widzimy, że w zbiorze danych występuje znacznie większa liczba osób zarabiających poniżej 50 tys. \$ (lewy słupek) od osób zarabiających powyżej 50 tys. \$ (prawy słupek). W trakcie analizy określimy podziały drzewa, które pozwolą wyodrębnić grupy w jak największym stopniu jednorodne ze względu na poziom dochodu. W pierwszej kolejności musimy wybrać zmienną, której użyjemy do pierwszego podziału. By móc to zrobić, warto wspomóc się rankingiem predyktorów dostępnym po naciśnięciu przycisku *Stat. Predyktorów*.

	Typ podziału	Poprawa
Związek	Automatycznie	0,074
Stan cywilny	Automatycznie	0,069
Liczba lat kształcenia	Automatycznie	0,039
Zawód	Automatycznie	0,035
Wykształcenie	Automatycznie	0,035
Wiek	Automatycznie	0,030
Płeć	Automatycznie	0,017
Grupa zawodowa	Automatycznie	0,008
Rasa	Automatycznie	0,004
Kraj pochodzenia	Automatycznie	0,003

Widzimy, że analizowaną grupę najsilniej ze względu na *Dochód* różnicuje zmienna *Związek* i właśnie ta zmienna posłuży nam do wykonania pierwszego podziału.



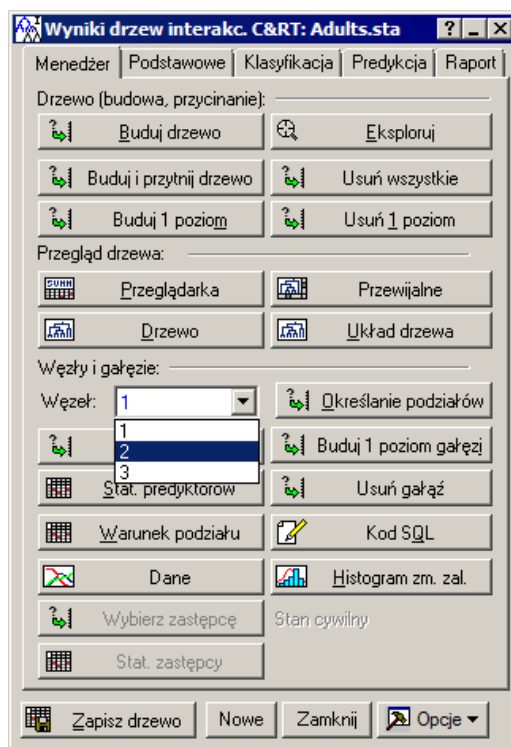
Klikamy przycisk *Określanie podziałów* i w wyświetlonym oknie widzimy, że algorytm sugeruje nam wprowadzenie podziału dla zmiennej *Związek* w ten sposób, by w lewym węźle znalazły się osoby zameżne (*Żona*, *Mąż*), natomiast w prawym węźle pozostałe osoby. Akceptujemy tę propozycję podziału, klikając *Buduj*, co pozwoli wprowadzić nasz podział do modelu drzewa.



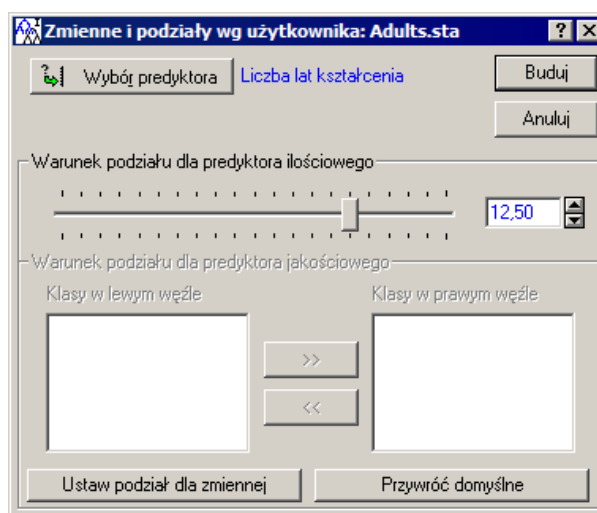
Po wprowadzeniu podziału widzimy, że powstałe dwa węzły znacznie różnią się między sobą, jeśli chodzi o rozkład zmiennej *Dochód*. Prawie połowa osób z lewego węzła osiąga dochód powyżej 50 tys. \$. W węźle prawym z kolei zdecydowaną przewagę mają osoby z dochodem poniżej 50 tys. \$.



Oczywiście powstałe węzły drzewa (liście) mogą podlegać dalszemu podziałowi. W kolejnym kroku określimy optymalny podział dla lewego węzła ( $ID=2$ ). W tym celu na liście rozwijalnej *Węzeł* wybieramy pozycję 2, a następnie klikamy przycisk *Określanie Podziałów*.



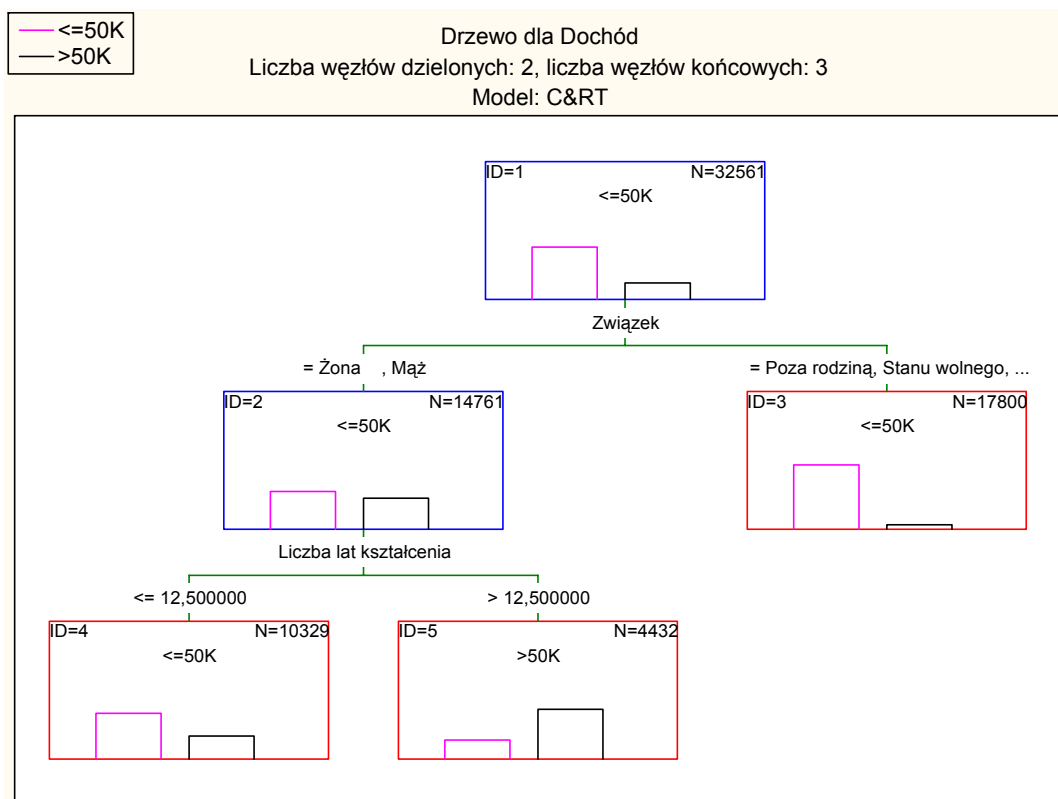
Tym razem najlepszą zmienną będącą podstawą podziału jest zmienna *Liczba lat kształcenia*. Program proponuje określić punkt podziału na poziomie 12,5, akceptujemy ten punkt podziału, klikając *Buduj*.<sup>10</sup>



<sup>10</sup> Korzystając z wiedzy eksperckiej, moglibyśmy zmienić zarówno zmienną służącą do podziału, jak też i poziom wprowadzanego podziału.



W wyniku podziału otrzymaliśmy nowe drzewo, które podzieliło zbiór danych na trzy segmenty. Jeśli dana osoba jest żonata lub zamężna oraz kształciła się dłużej niż 12,5 roku wtedy trafia do węzła 5 – możemy ją zaliczyć do grona osób o najwyższych dochodach. W węźle 4 znajdują się osoby wykazujące w przeważającej mierze dochody poniżej 50 tys. \$, choć segment ten zawiera też spory odsetek osób z wyższymi dochodami, natomiast węzeł 3 to segment osób o najniższych dochodach.



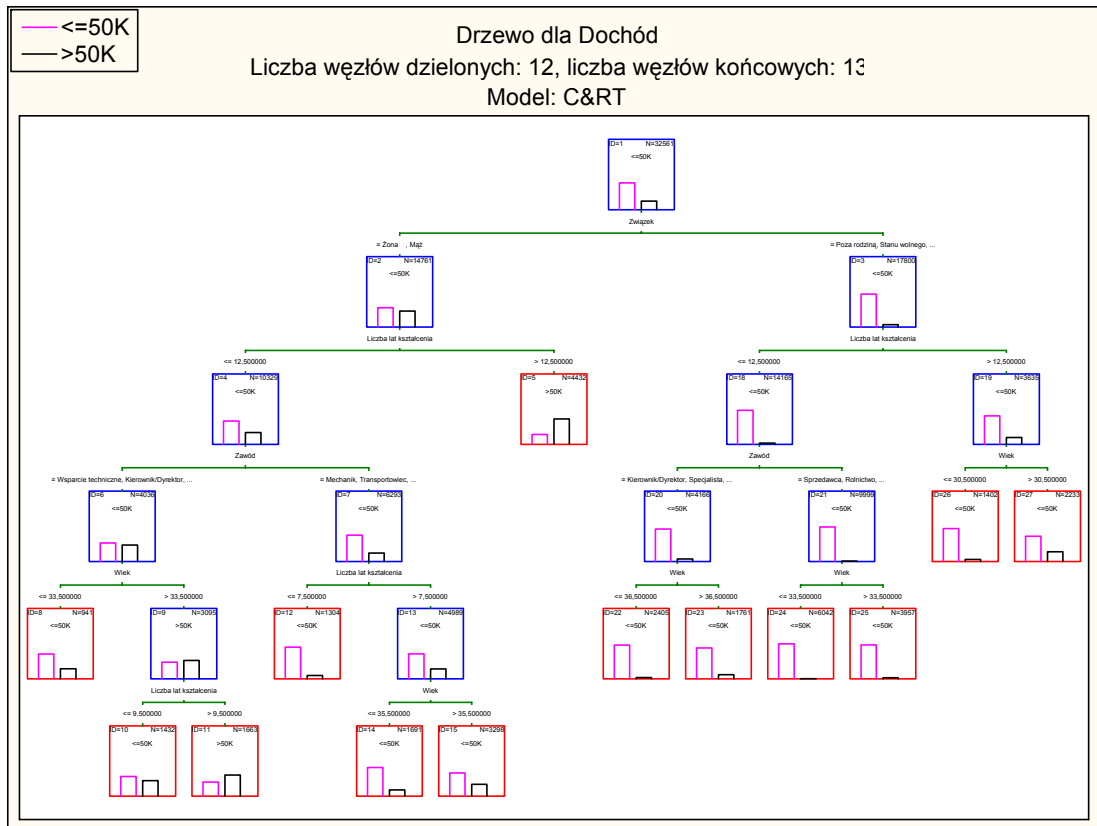
Jeśli uzyskane wyniki segmentacji byłyby dla nas zbyt ogólne, każdy z utworzonych węzłów końcowych możemy dalej dzielić w celu określenia bardziej jednorodnych segmentów. Poza zaprezentowanym powyżej ręcznym budowaniem drzewa zawsze mamy możliwość zbudowania go w sposób automatyczny. Po naciśnięciu przycisku *Buduj drzewo* otrzymujemy zbudowany automatycznie dużo bardziej skomplikowany model składający się z 13 węzłów końcowych (zob. rys. poniżej).

Dla wszystkich utworzonych węzłów końcowych możemy podać łatwą w interpretacji regułę opisującą osoby przypisane do poszczególnych grup ryzyka (podobnie jak uczyniliśmy to z węzłem 5).

Istotnym problemem, jaki możemy napotkać podczas budowy modelu, jest przeuczenie (nadmierne dopasowanie do danych). Niektóre z zaproponowanych przez automat podziałów tworzą węzły o bardzo małej licznosci, przez co wzrasta ryzyko, że reguły przez nie opisywane są jedynie wynikiem szumu zawartego w danych. Aby uniknąć budowy modelu nadmiernie dopasowanego do danych, możemy zaproponowany przez automat model uprościć, przycinając gałęzie, które mają zbyt małą licznosc bądź zbudować model



za pomocą *V-krotnego sprawdzianu krzyżowego*, dzięki czemu algorytm automatycznie określi optymalną głębokość drzewa.



## Literatura

1. Sagan A., Łapczyński M., *Techniki segmentacji w badaniach rynkowych*, Materiały szkoleniowe StatSoft Polska, 2009.
2. Sokołowski A. *Empiryczne testy istotności w taksonomii*, Akademia Ekonomiczna w Krakowie, Zeszyty Naukowe, Kraków 1992
3. *STATISTICA Neural Networks PL. Wprowadzenie do sieci neuronowych*, StatSoft Polska, 2001.