Obiekty danych

- Zbiory danych składają się z obiektów.
  - O**biekt danych** reprezentuje pojęcie.
  - przykłady:
    - baza sprzedażowa: klienci, sprzedaż, szczegóły sprzedaży
    - medyczna baza: pacjenci, terapie
    - baza uniwersytecka: studenci, wykładowcy, kursy
  - Zwane również *próbkami , przykładami, instancjami, punktami danych, krotkami*.
- Obiekty danych są opisane **atrybutami**.
- Rekordy bazy-> obiekty danych; kolumny ->atrybuty.

Atrybuty

- **Atrybuty (**lub **wymiary: DW, cechy: ML, zmienne: ST**): pola danych, reprezentujące charakterystykę lub cechę obiektu danych.
  - *np. klient _ID, nazwa, adres*
- Zbiór atrybutów używany do opisania danego obiektu – *wektor atrybutów* (lub *wektor cech*)
- Typy atrybutu określane przez zbiór możliwych wartości:
  - Nominalne
  - Binarne
  - Porządkowe
  - Numeryczne
    - przedziałowe (interval scaled)
    - ilorazowe (ratio-scaled)

Typy atrybutów

- **Nominalne (kategoryczne,wyliczeniowe):** kategorie, stany, “nazwy rzeczy”
  - *Kolor_włosów = {czarny, brązowy, blond, kasztanowy, siwy}*
  - stan cywilny, zawód, ale również np. identyfikatory, kody pocztowe
  - jakościowa klasyfikacja, brak porządku, można reprezentować numerycznie
- **Binarne**
  - Nominalne z 2 stanami (0 lub 1)
    - np. palacz
  - Symetryczne: obydwa stany jednakowo ważne
    - np. płeć (?)
  - Asymetryczne: wyniki nie są tak samo ważne
    - np. testy medyczne (pozytywne - negatywne)
    - konwencja: przypisz 1 do najważniejszych wyników (np. HIV pozytywny)
- **Porządkowe (jakościowe)**
  - Wartości mają znaczący porządek (ranking) ale odległości m. kolejnymi wartościami nieznane.
    - można je uzyskać również poprzez dyskretyzację
  - np. wielkość = {mały, średni, duży}

Numeryczne typy atrybutów

- są ilościowe, mierzalne (wartości całkowite lub rzeczywiste)
- **Przedziałowe** *(interval-scaled)*
  - pozwalają rangować mierzone obiekty i mierzyć różnice m. nimi
  - brak punktu absolutnego zera
    - np. *temperatura w C˚lub F˚, daty kalendarzowe*
- **Ilorazowe** *(ratio-scaled)*
  - istnieje **punkt absolutny zera skali**
  - stosunki między wartościami mogą być zdefiniowane w sposób znaczący (10 K˚ jest dwa razy większa niż 5 K˚).
    - np. *temperatura w Kelvinach, przestrzeń, czas*

Dyskretne a ciągłe atrybuty

- **Inny podział (ML)**
- **Atrybut dyskretny**
  - Ma skończony lub policzalny zbiór wartości
    - np. kody pocztowe, zawody, zbiór słów w kolekcji dokumentów
  - Można reprezentować jako wartości całkowite
  - Atrybuty binarne: szczególny przypadek dyskretnych
- **Atrybut ciągły**
  - Liczy rzeczywiste wartości atrybutu
    - np. temperatura, wysokość, waga
  - W praktyce, rzeczywiste wartości mogą być mierzone i reprezentowane przy użyciu skończonego zbioru liczb

SPSS Modeler

Węzeł TYPY

Węzeł TYPY

- Odpowiada za zarządzanie opisem danych w strumieniu analitycznym – jeden z najważniejszych węzłów
  - dodatkowo spełnia wiele innych funkcji:
    - etykietowanie
    - możliwość definiowania braków danych użytkownika
- Przycisk "Odczytaj wartości"
  - zaprezentowane zostają zmienne obecne w zbiorze wejściowym oraz ich typ
  - kolumna "Wartości" – można obserwować jakie wartości przyjmują poszczególne zmienne

POZIOM POMIARU

- **Domyślny** – program sam rozpoznaje z jakiego typu danymi ma do czynienia – ustawienie zostanie zmienione po pierwszym odczytaniu wartości zmiennej w węźle
- **Ilościowa** – zmienna przybiera wartości będące liczbami całkowitymi, rzeczywistymi lub zmiennymi typami data lub czas
- **Jakościowa** – pojawia się dla zmiennych tekstowych, dla których liczba unikalnych wartości (kategorii) nie jest znana (nie zostały zeskanowane wartości zmiennej)
- **Flaga** – zmienna przyjmuje 2 wykluczające się wartości: prawda/fałsz, 0/1, tak/nie
- **Nominalna** – przyjmuje wiele wartości, które mogą być wyrażone tekstowo lub liczbowo choć zapisane w postaci symboli liczbowych, nie podlegają regułom operacji matematycznych (np. wartości zmiennej "region": 1, 2, 3 itd. nie dają się sumować)
- **Porządkowa** – zmienna przyjmuje kilka wartości zwanych kategoriami, które posiadają określony porządek, mogą to być zarówno wartości liczbowe jak i tekstowe (np. rozmiar "mały, "średni", "duży" lub ocena szkolna – można powiedzieć, że 5 to ocena wyższa niż 4, a "mały" to mniej niż "średni")
- **Nieokreślony** – zmienna, której wartości nie powinny być skanowane (np. dla identyfikatora klienta, czy nr rachunku, które nie będą wykorzystywane w analizie

BRAKI

- Możliwość zdefiniowania braków danych dla poszczególnych zmiennych
  - obok systemowych braków danych można zdefiniować braki danych użytkownika, np. 9 może być traktowana jako "nie dotyczy"
  -

SPRAWDŹ

- Decydujemy jak program ma się zachować w sytuacji gdy natrafi na braki danych lub gdy wystąpi wartość spoza zakresu lub zdefiniowanej listy
- **Brak** – sprawdzanie jest wyłączone dla danej zmiennej
- **Wyzeruj** – zastąpienie systemowym brakiem danych ($null$)
- **Wymuś** – dla zmiennych typu:
  - flaga – niepoprawny wpis jest zastępowany wartością fałsz
  - jakościowa (nominalna lub porządkowa) – nieznana wartość jest zmieniona na pierwszą wartość ze zbioru wartości
  - ilościowa – wartość która jest ponad górną granicą jest sprowadzana do postaci najwyższej dopuszczalnej wartości; wartość poniżej najniższej analogicznie do najniższej; w sytuacji gdy wystąpi brak danych przypisywana jest wartość środkowa
- **Odrzuć** – gdy niewłaściwy wpis zostanie wykryty rekord jest automatycznie usuwany
- **Ostrzeż** – w komunikatach wyświetlany jest raport o wystąpieniu rekordów nieprawidłowych
- **Przerwij** – gdy program napotka na wartość spoza zakresu automatycznie przerywa dalsze przetwarzanie danych

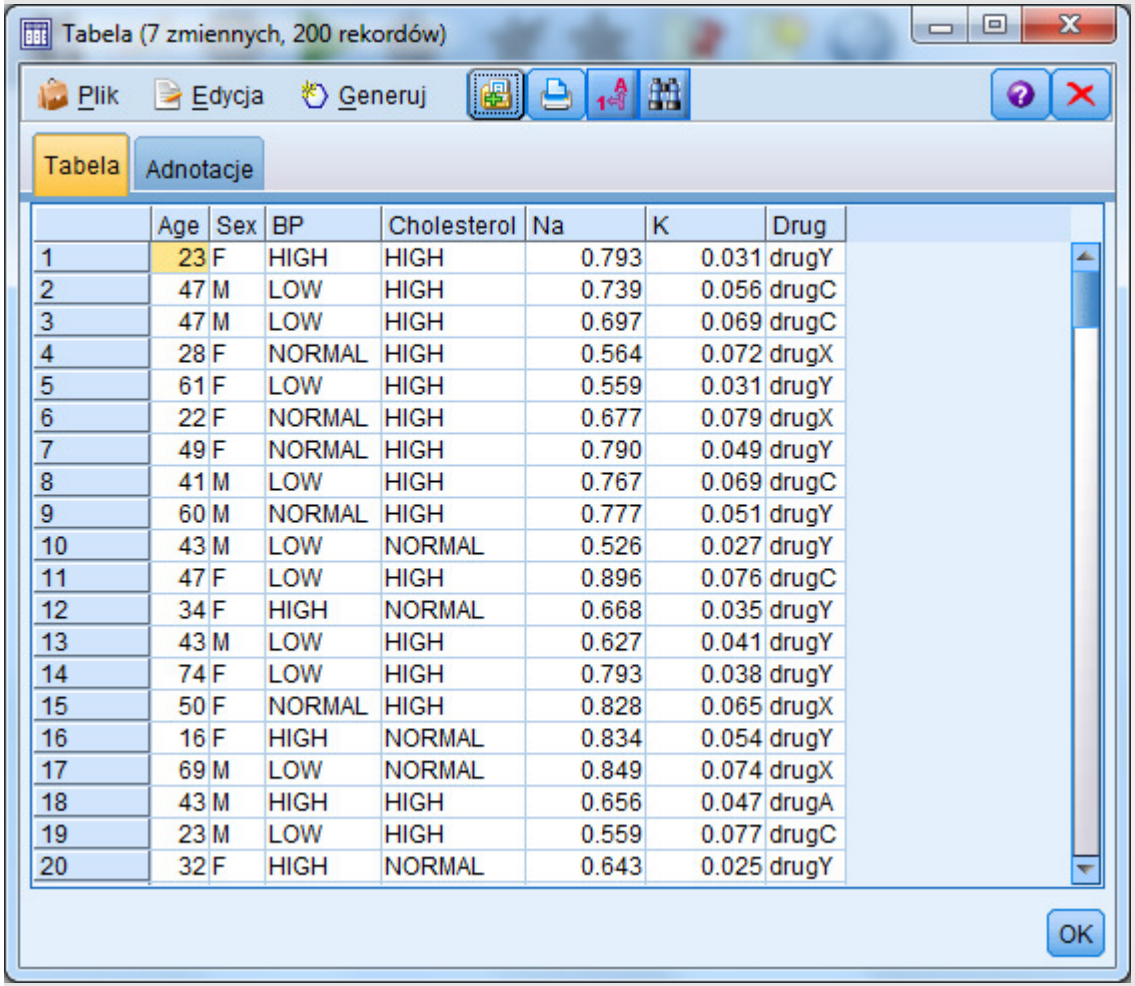
ROLA

- Ustalenie która zmienna ma być zmienną zależną ("**Przewidywana**") a która niezależną/predyktorem ("**Wejściowa**")
- Gdy zmienne może być zarówno przewidywaną lub predyktorem (np. w regułach asocjacyjnych) wtedy zaznacza się Opcję "**Obydwie**"
- **Podział** – wskazanie zmiennej dzielącej zbiór danych na podzbiory: testowy i uczący (częściej węzeł "Podział")
- **Separacja** (dla zmiennych nominalnych, porządkowych, flag) – na etapie modelowania modele mają być budowane oddzielnie dla każdej wartości przyjmowanych przez tę zmienną
- **Ważenie rekordów** (dla C&RT, CHAID, QUEST) – użycie wartości tej zmiennej jako czynnika ważenia liczebności rekordów
- **ID rekordów** – używana tylko przez modele liniowe
- **Brak** – zmienna w modelu nie będzie wykorzystywana
- Dwa widoki: "widok aktualnych zmiennych", "widok ustawień niewykorzystanych zmiennych"

Zakładka FORMAT

- Zmienna – lista zmiennych znajdujących się w zbiorze danych
- Format – można dokonać zmiany ustawień dotyczących formatu daty, czasu, liczb, separatorów dziesiętnych, szerokości kolumn, wyrównania, etc.
- Wyrównanie – określa sposób prezentacji danch np. w tabelach
- Szerokość kolumn – określa domyślną szerokość prezentowanej kolumny

Ćwiczenie 1 – przewidywanie efektywności leczenia - dane



	Age	Sex	BP	Cholesterol	Na	K	Drug
1	23	F	HIGH	HIGH	0.793	0.031	drugY
2	47	M	LOW	HIGH	0.739	0.056	drugC
3	47	M	LOW	HIGH	0.697	0.069	drugC
4	28	F	NORMAL	HIGH	0.564	0.072	drugX
5	61	F	LOW	HIGH	0.559	0.031	drugY
6	22	F	NORMAL	HIGH	0.677	0.079	drugX
7	49	F	NORMAL	HIGH	0.790	0.049	drugY
8	41	M	LOW	HIGH	0.767	0.069	drugC
9	60	M	NORMAL	HIGH	0.777	0.051	drugY
10	43	M	LOW	NORMAL	0.526	0.027	drugY
11	47	F	LOW	HIGH	0.896	0.076	drugC
12	34	F	HIGH	NORMAL	0.668	0.035	drugY
13	43	M	LOW	HIGH	0.627	0.041	drugY
14	74	F	LOW	HIGH	0.793	0.038	drugY
15	50	F	NORMAL	HIGH	0.828	0.065	drugX
16	16	F	HIGH	NORMAL	0.834	0.054	drugY
17	69	M	LOW	NORMAL	0.849	0.074	drugX
18	43	M	HIGH	HIGH	0.656	0.047	drugA
19	23	M	LOW	HIGH	0.559	0.077	drugC
20	32	F	HIGH	NORMAL	0.643	0.025	drugY

Age	Wiek (ilościowa)
Sex	Płeć (M lub F)
BP	Ciśnienie krwi: HIGH, NORMAL lub LOW
Cholesterol	Poziom cholesterolu: NORMAL lub HIGH
Na	Poziom sodu (ilościowa)
K	Poziom potasu (ilościowa)
Drug	Lek, który podziałał na pacjenta (zm. doc.)

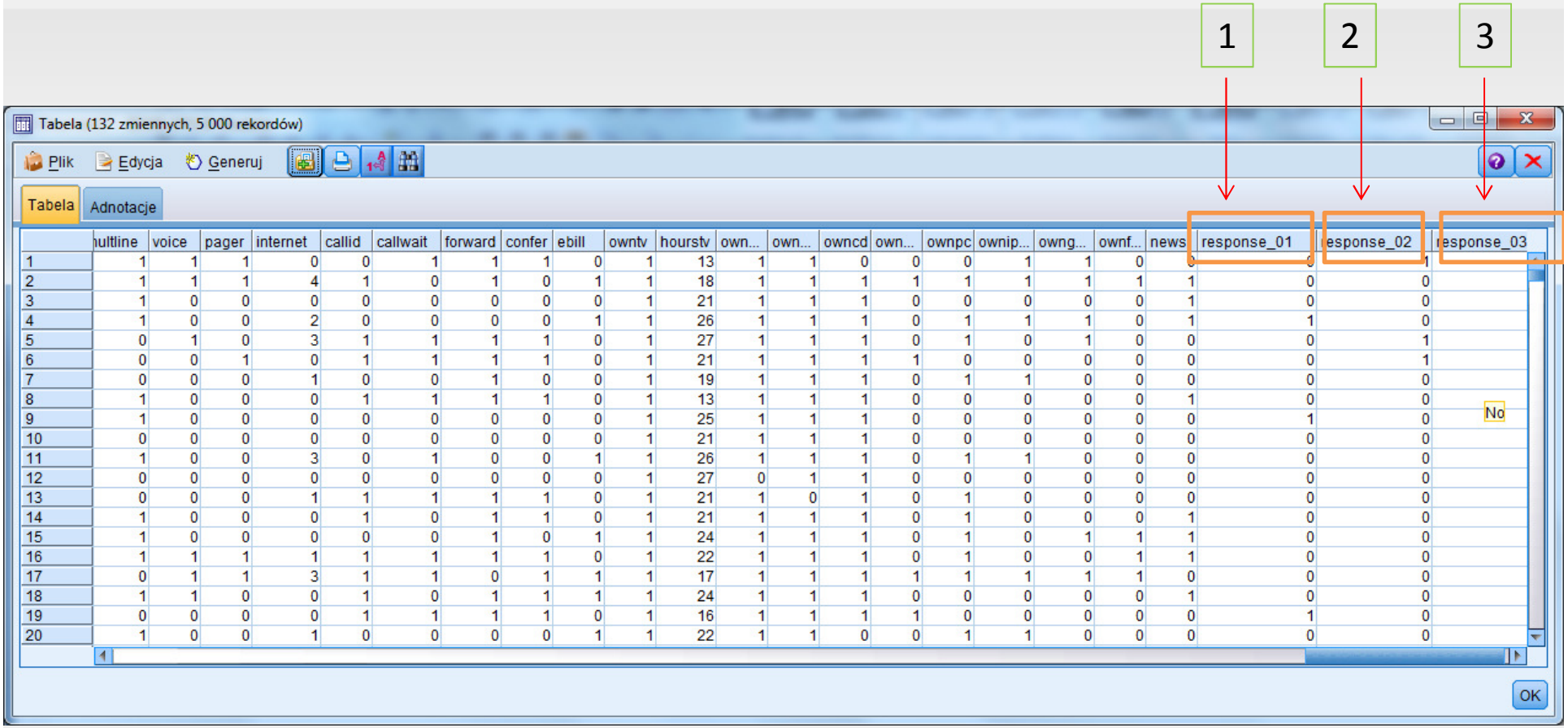
Ćwiczenie 1 – przewidywanie efektywności leczenia

- Jaki czynnik ma największy wpływ na wybór terapii?
  - jaka część pacjentów odpowiedziała pozytywnie na terapię danym lekiem?
    - węzeł "Rozkładu" – dostajemy informację, że pacjenci odpowiadali pozytywnie najczęściej na lek Y a najrzadziej na B i C
  - jakie czynnik wpływają na zmienną docelową "Drug"
    - wiedza dziedzinowa: koncentracja sodu i potasy we krwi są istotnymi czynnikami
    - wykres "Rozrzutu" – pokazuje progi powyżej których poprawnym lekiem jest Y, a poniżej nigdy (próg ten to stosunek sodu i potasu)
  - wizualizacja zależności między różnymi kategoriami
    - wykres "Sieciowy" – można zaobserwować tylko "drugY" jest związany ze wszystkimi trzema poziomami ciśnienia we krwi
  - ponieważ stosunek sodu do potasu wydaje się być czynnikiem wpływającym na zastosowanie leku Y – tworzymy nowe pole – węzeł "Wyliczanie"

Ćwiczenie 1 – przewidywanie efektywności leczenia – c.d.

- Analizując dane można stawiać różne hipotezy – nie można jednak w pełni określić wszystkich zależności
- Budujemy model klasyfikacyjny C5.0
  - usuwamy pola Na i K (bo mamy nowe) – węzeł "Filtrowanie"
  - ustawiamy zmienną "Drug" jako "Przewidywana" – węzeł "Typy"
  - węzeł "C5.0" z zakładki "Modelowanie"
  - wygenerowany model pojawi się w prawym górnym rogu
- Szacowanie dokładności
  - węzeł "Analiza" z zakładki "Wyniki"

Ćwiczenie 2 – badanie predyktorów - dane



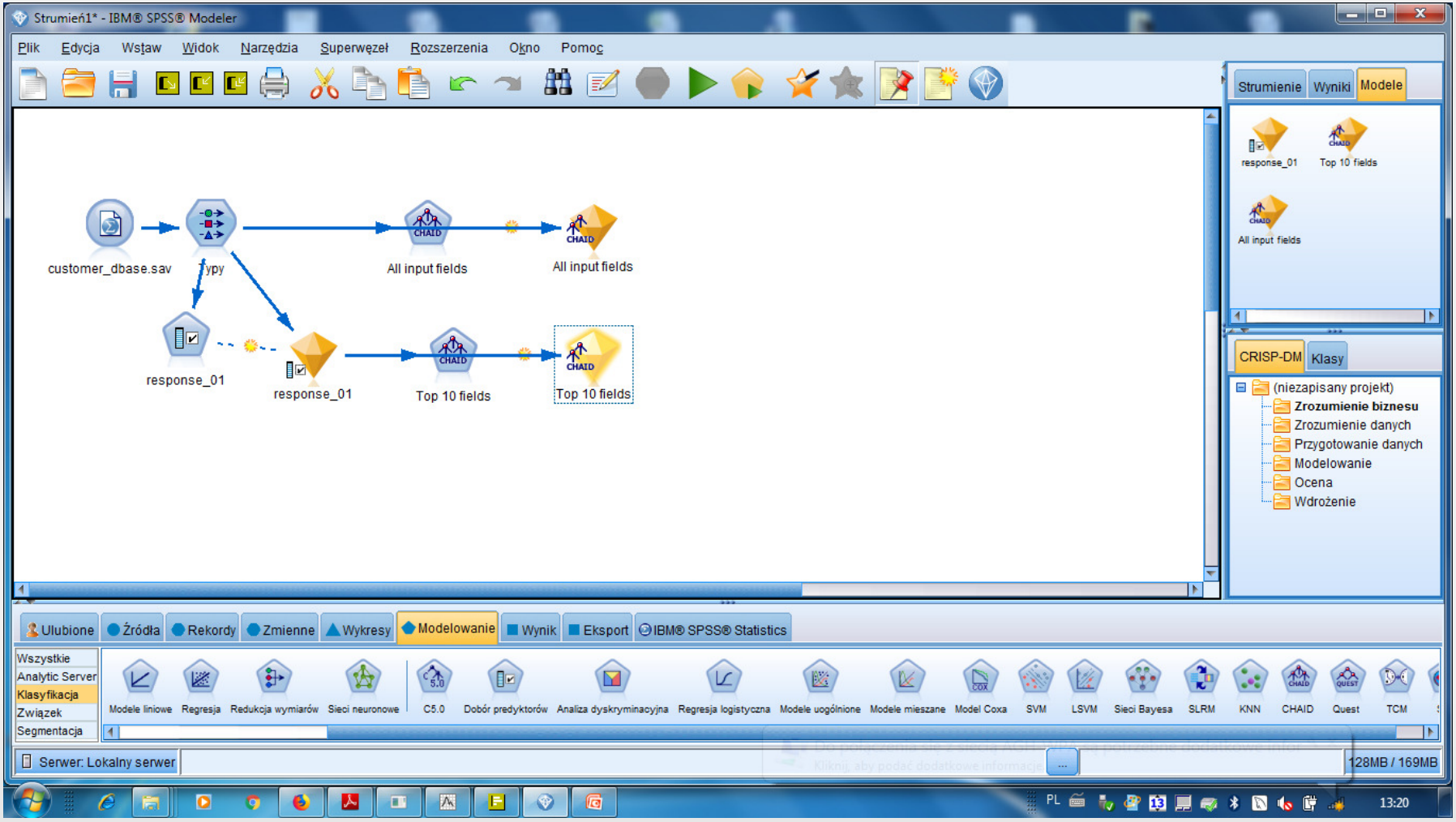
Ćwiczenie 2 – badanie predyktorów

- Cel: identyfikacja atrybutów, które mają największy wpływ na predykcję atrybutów wyjściowych
- 3 zmienne docelowe – czy klient odpowiedział na dane zapytanie ofertowe (3)
- Wspomaganie predykcji którzy klienci odpowiedzą w przyszłości na podobne oferty
  - za każdym razem jedna oferta wybierana jest jako cel
  - model drzewa CHAID – bez wyboru cech i z wyborem cech

Ćwiczenie 2 – budowanie strumienia

- węzeł "Typy" – ustalenie roli zmiennych
  - "response_01" = "Przewidywana"
  - "custid", "response_02", "response_03" = "Brak"
- węzeł "Wybór predyktorów" z zakładki "Modelowanie" – do ustalenia ważności predyktorów
- wstawienie modelu do strumienia (o ile nie jest wstawiony) za węzłem "Typy"
  - wybór 10 pierwszych predyktorów
- porównanie z wyborem predyktorów i bez:
  - jeden węzeł "CHAID" połączyć z "Typy" (bez)
  - drugi węzeł "CHAID" połączyć z wygenerowanym modelem
- porównać czas budowy modelu i wielkość drzewa
- zrobić dla pozostałych ofert

Wynikowy strumień



Ćwiczenie 3 – tworzenie profili klientów - dane

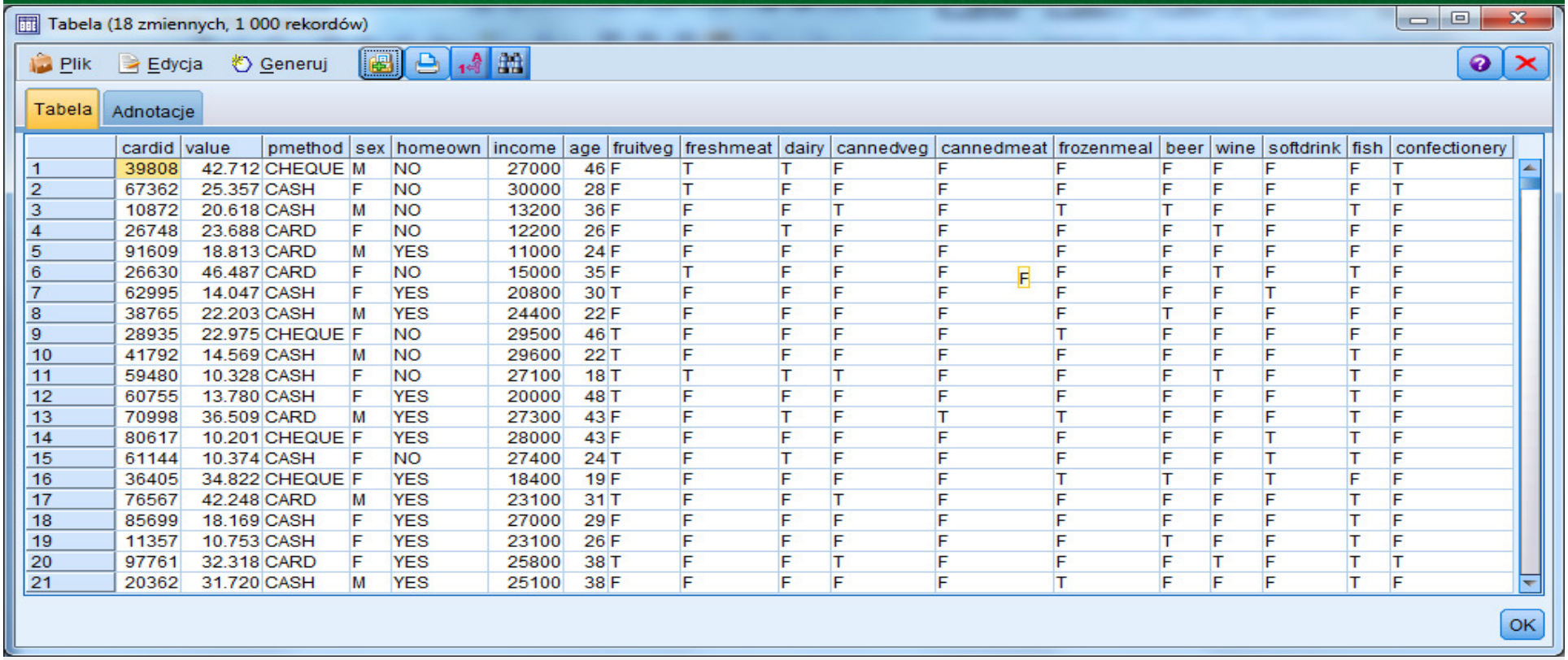


Tabela (18 zmiennych, 1 000 rekordów)

Plik Edycja Generuj

Tabela Adnotacje

	cardid	value	pmethod	sex	homeown	income	age	fruitveg	freshmeat	dairy	cannedveg	cannedmeat	frozenmeal	beer	wine	softdrink	fish	confectionery
1	39808	42.712	CHEQUE	M	NO	27000	46	F	T	T	F	F	F	F	F	F	F	T
2	67362	25.357	CASH	F	NO	30000	28	F	T	F	F	F	F	F	F	F	F	T
3	10872	20.618	CASH	M	NO	13200	36	F	F	F	T	F	T	T	F	F	T	F
4	26748	23.688	CARD	F	NO	12200	26	F	F	T	F	F	F	F	T	F	F	F
5	91609	18.813	CARD	M	YES	11000	24	F	F	F	F	F	F	F	F	F	F	F
6	26630	46.487	CARD	F	NO	15000	35	F	T	F	F	F	F	F	T	F	T	F
7	62995	14.047	CASH	F	YES	20800	30	T	F	F	F	F	F	F	F	T	F	F
8	38765	22.203	CASH	M	YES	24400	22	F	F	F	F	F	F	T	F	F	F	F
9	28935	22.975	CHEQUE	F	NO	29500	46	T	F	F	F	F	T	F	F	F	F	F
10	41792	14.569	CASH	M	NO	29600	22	T	F	F	F	F	F	F	F	F	T	F
11	59480	10.328	CASH	F	NO	27100	18	T	T	T	T	F	F	F	T	F	T	F
12	60755	13.780	CASH	F	YES	20000	48	T	F	F	F	F	F	F	F	F	T	F
13	70998	36.509	CARD	M	YES	27300	43	F	F	T	F	T	T	F	F	F	T	F
14	80617	10.201	CHEQUE	F	YES	28000	43	F	F	F	F	F	F	F	F	T	T	F
15	61144	10.374	CASH	F	NO	27400	24	T	F	T	F	F	F	F	F	T	T	F
16	36405	34.822	CHEQUE	F	YES	18400	19	F	F	F	F	F	T	T	F	T	F	F
17	76567	42.248	CARD	M	YES	23100	31	T	F	F	T	F	F	F	F	F	T	F
18	85699	18.169	CASH	F	YES	27000	29	F	F	F	F	F	F	F	F	F	T	F
19	11357	10.753	CASH	F	YES	23100	26	F	F	F	F	F	F	T	F	F	T	F
20	97761	32.318	CARD	F	YES	25800	38	T	F	F	T	F	F	F	T	F	T	T
21	20362	31.720	CASH	M	YES	25100	38	F	F	F	F	F	T	F	F	F	T	F

OK

cardid – nr karty lojalnościowej
value – wartość koszyka
pmethod – sposób płatności

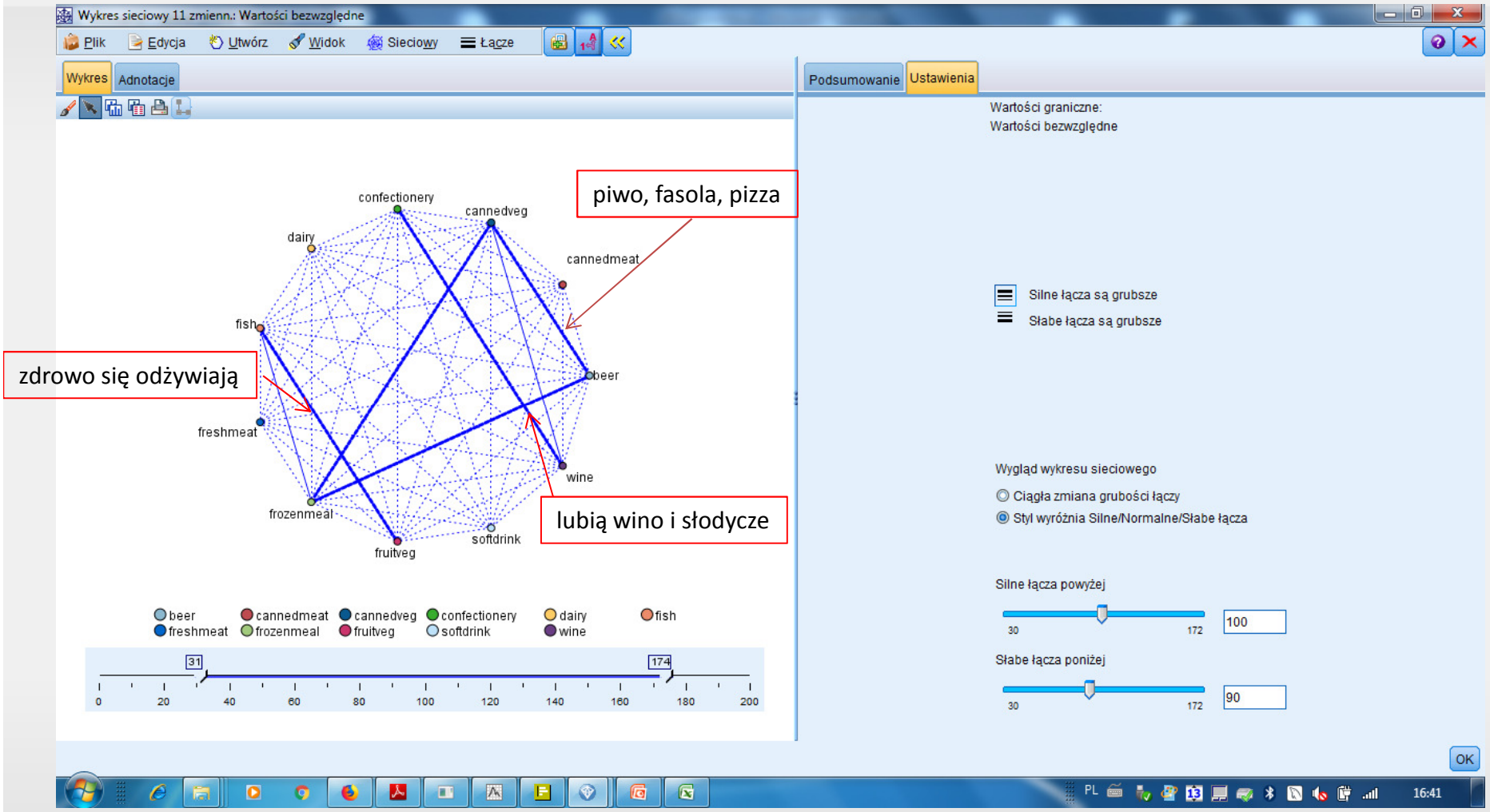
Charakterystyka klienta:
- sex - płeć
- homeown - czy klient posiada dom
- income - dochód
- age - wiek

Zawartość koszyka
flaga określająca czy produkt z danej kategorii znalazł się w koszyku
fruitveg, freshmeat, dairy, cannedveg, cannedmeat, frozenmeal, beer, wine, softdrink, fish, confectionery

Ćwiczenie 3 – tworzenie profili klientów

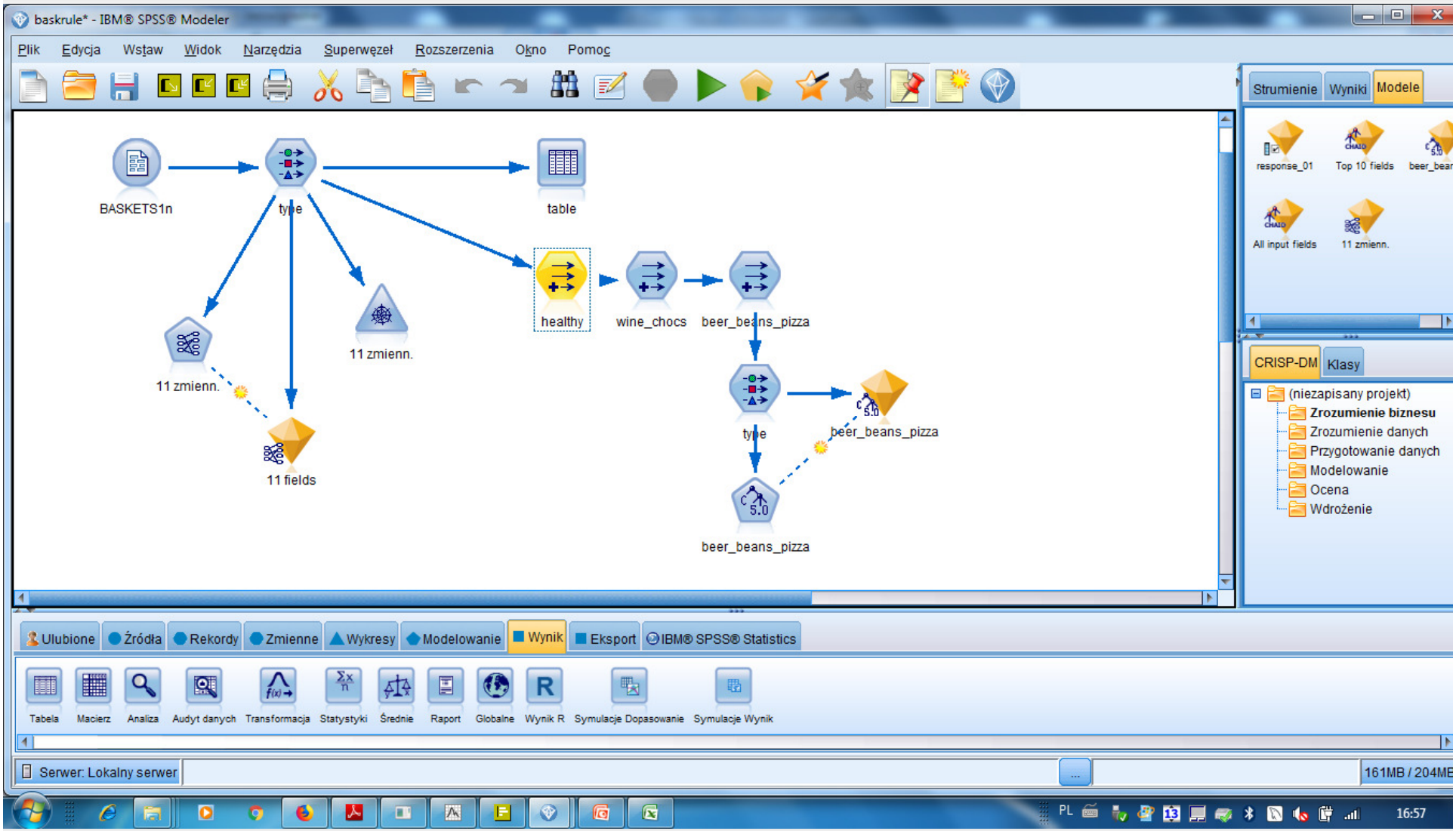
- Cel: znalezienie zależności w rodzaju: jeżeli dużo zarabia i jest w średnim wieku to kupuje zdrową żywność
- Dwa kroki:
  1. analiza koszyka zakupowego
  2. dołączenie danych o kliencie i uzyskanie informacji jacy klienci kupują dane grupy produktów
- Ad1.
  - wybór tylko produktów ("Obydwie"), pozostałe na "Brak"
  - węzeł "Apriori"
  - podłączyć węzeł "Sieciowy" – do oznaczenia grup

3 grupy klientów



Ćwiczenie 3 – tworzenie profili klientów

- Ad2
  - ustawmy flagę dla każdej znalezionej grupy
  - oznaczmy każdego klienta flagą przynależności do każdej grupy
  - węzeł C5.0 do indukcji reguł dla każdej z flag


baskrule* - IBM® SPSS® Modeler
Plik Edycja Wstaw Widok Narzędzia Superwęzeł Rozszerzenia Okno Pomoc
Strumienie Wyniki Modele
response_01
Top 10 fields
All input fields
11 zmienn.
CRISP-DM Klasy
(niezapisany projekt)
Zrozumienie biznesu
Zrozumienie danych
Przygotowanie danych
Modelowanie
Ocena
Wdrożenie
BASKETS1n
type
table
healthy
wine_chocs
beer_beans_pizza
11 zmienn.
11 zmienn.
11 fields
type
beer_beans_pizza
beer_beans_pizza
Ulubione Źródła Rekordy Zmienne Wykresy Modelowanie Wynik Eksport IBM® SPSS® Statistics
Tabela Macierz Analiza Audyt danych Transformacja Statystyki Średnie Raport Globalne Wynik R Symulacje Dopasowanie Symulacje Wynik
Serwer: Lokalny serwer
161MB / 204ME
PL
16:57