DATA MINING I JEGO REALIZACJA W *STATISTICA DATA MINER*

dr Janusz Wątroba, mgr inż. Tomasz Kowalski, mgr Tomasz Demski, StatSoft Polska

Wstęp

Wraz z rozwojem technologii informatycznej pojawiły się nowe możliwości pod względem ilości gromadzonych danych oraz szybkości ich przetwarzania. W tej chwili można śmiało powiedzieć, że generowanie, przechowywanie czy przesyłanie informacji nie stanowi żadnego problemu. Kluczowym zagadnieniem jest natomiast wydobycie z danych użytecznych informacji, które pomogą podjąć decyzję szybciej i trafniej niż konkurencja. W odpowiedzi na takie możliwości, a jednocześnie nowe wymagania powstały narzędzia *data mining* (zgłębiania danych). Można jednak zapytać, które z nich rzeczywiście służą do celów *data mining* i spełniają wymagania tej metodologii, a które są tylko cząstkowym rozwiązaniem?

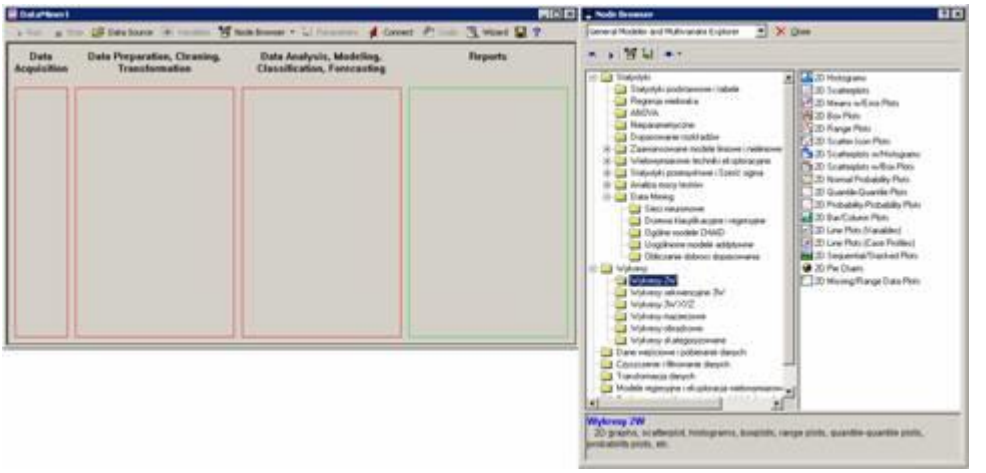
Z praktycznego punktu widzenia narzędzie *data mining* powinno umożliwiać:

- Wydajny dostęp do danych
- Przygotowanie danych dla potrzeb analiz typu *data mining*
- Przeprowadzenie analiz typu *data mining* (nawet dla ogromnych zbiorów danych)
- Wizualizację i raportowanie wyników analiz

Istnieje szereg narzędzi umożliwiających wykonanie powyższych etapów *data mining*. Jedne z nich umożliwiają wydobycie danych z bazy danych, inne wykonywanie analiz na tych danych, a jeszcze inne generowanie raportów i podsumowań. Zazwyczaj programy te pochodzą od różnych producentów i różnią się nie tylko funkcjonalnością, ale również interfejsem użytkownika, a nawet niezgodnością formatu danych (co z góry dyskwalifikuje je nawet jako składnik systemu *data mining*). Dodatkową trudnością jest fakt, że każdego z tych programów trzeba uczyć się oddzielnie a ich współpraca może być przyczyną dodatkowych trudności [1].

System analizy danych, który nie jest w stanie operować na bardzo dużych zbiorach danych powinien zostać zaklasyfikowany do grupy narzędzi analitycznych, eksperymentalnych lub uczenia maszynowego. Natomiast system, który umożliwia jedynie przeszukiwanie dużych zbiorów danych i obliczanie wartości zagregowanych, lub tworzenie i wykonywanie zapytań do zewnętrznych baz danych jest systemem bazodanowym lub informacyjnym. Żaden z nich nie jest systemem *data mining* (zob. [3]).

STATISTICA Data Miner (*www.statsoft.pl/dataminer.html*) to kompleksowy, zintegrowany system *data mining*, który spełnia wymienione wyżej wymagania, a ponadto ma wiele unikalnych zalet.



Przestrzeń robocza *STATISTICA Data Miner* wraz z Przeglądarką węzłów

Interfejs *STATISTICA Data Miner* umożliwia proste i intuicyjne wykonywanie wszystkich zadań *data mining*. Przestrzeń robocza systemu została podzielona na cztery części, z których każda dotyczy konkretnych zadań *data mining*: dostępu do danych, przekształcania danych dla potrzeb analiz typu *data mining*, przeprowadzenia samej analizy oraz wizualizacji wyników i raportowania. Setki procedur analitycznych i wizualizacyjnych dostępnych jest w postaci węzłów analitycznych znajdujących się w przeglądarce węzłów. Po wstawieniu ich do przestrzeni roboczej *STATISTICA Data Miner* i połączeniu ze sobą otrzymamy kompletny projekt *data mining*. Dzięki takiemu trybowi pracy nawet skomplikowaną wieloetapową analizę możemy łatwo zbudować i modyfikować przeciągając obiekty myszą. Ponadto łatwo jest zorientować się w strukturze projektu.

Wszystkie elementy środowisko systemu możemy dostosowywać do własnych potrzeby i upodobań. Można m.in. zbudować własną przeglądarkę węzłów, tworzyć własne węzły analityczne (korzystając z makr nagranych podczas interakcyjnego specyfikowania analiz i wykresów), a nawet uruchamiać system i sterować nim z innych aplikacji (np. z MS Excel).

Wersja klient-serwer *STATISTICA Data Miner*, wykorzystuje przetwarzanie rozproszone i architekturę wielowarstwową, co pomaga optymalnie wykonywać bardzo złożone zadania obliczeniowe. Technologia ta umożliwia szybkie wykonywanie nawet bardzo dużych projektów z pełnym wykorzystaniem wielu procesorów serwera lub wielu serwerów pracujących równolegle.

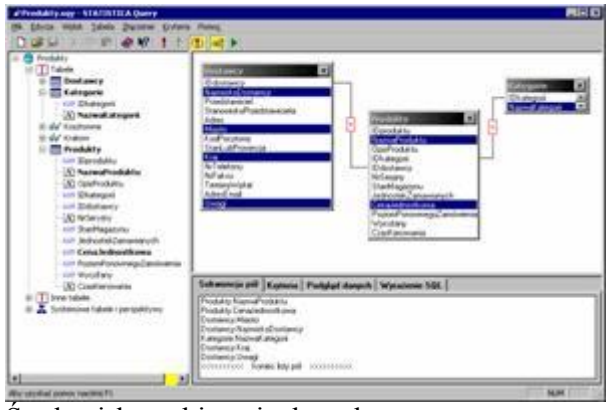
W przypadku wersji klient-serwer *STATISTICA Data Miner* wszystkie obliczenia odbywają się na serwerze (lub wielu serwerach) *STATISTICA*, a stacja robocza

użytkownika obsługuje wyłącznie interfejs programu. Na komputerze użytkownika nie trzeba instalować żadnego oprogramowania - wystarczy przeglądarka internetowa.

W dalszej części artykułu postaramy się bliżej scharakteryzować wymagania stawiane przed narzędziami *data mining* oraz przedstawić środowisko kompleksowego systemu *data mining STATISTICA Data Miner*.

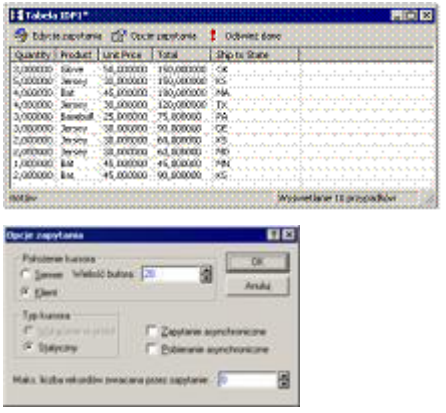
Pobieranie danych

Z założenia proces *data mining* dotyczy bardzo dużej ilości danych, przechowywanych w bazach i hurtowniach danych często o bardzo złożonej i skomplikowanej strukturze. Standardowym językiem pobierania danych z bazy danych jest SQL. Użytkownicy nie znający tego języka i zaawansowanej technologii informatycznej również potrzebują łatwego dostępu do danych gromadzonych w różnych repozytoriach danych. Każde narzędzie *data mining*powinno więc posiadać wbudowane mechanizmy dostępu do zewnętrznych baz danych gwarantujące użytkownikowi interakcyjną budowę zapytań do bazy danych i łatwy dostęp do danych.



Środowisko pobierania danych

- *STATISTICA Query*



Dane przeniesione do tabeli IDP oraz okno opcji umożliwiających wybór zdalnego przetwarzania danych

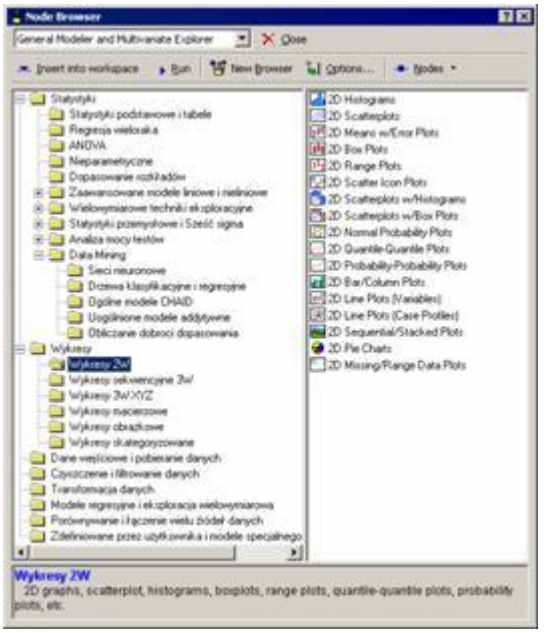
W systemie *STATISTICA Data Miner* możliwe jest nie tylko korzystanie z danych zapisanych lokalnie w pliku *STATISTICA*, ale również przetwarzanie zapytań po stronie serwera, to znaczy bez konieczności importowania danych i tworzenia pliku lokalnego. Technologia ta (*IDP - In-place Database Processing*), jest użyteczna przy przetwarzaniu bardzo dużych zbiorów danych; w takich przypadkach jej zastosowanie daje duży zysk wydajności i umożliwia przetwarzanie zbiorów danych o wielkości przekraczającej pojemność urządzeń lokalnych. W procesie pobierania danych z bazy danych można posłużyć się narzędziem *STATISTICA Query*, umożliwiającym interakcyjną budowę zapytania i pobranie lub połączenie w ten sposób danych ze środowiskiem *STATISTICA Data Miner*.

Projekt *data mining* może być automatycznie aktualizowany przy każdej zmianie danych. Dzięki temu możemy np. zbudować system automatycznie wykrywający podejrzane transakcje i powiadamiający o tym odpowiednie osoby.

Przekształcanie danych dla potrzeb analiz

Często dane w bazie danych gromadzone są automatycznie i dotyczą wszystkich gałęzi przedsiębiorstwa. Czasem nie potrzebujemy aż tak różnorodnych danych, ale tylko ściśle określonych cech. Po połączeniu systemu *data mining* z wybranymi danymi może okazać się, że konieczne jest wstępne przetworzenie danych dla potrzeb konkretnych analiz typu *data mining*.

Na przykład, gdy w modelu *data mining* korzystamy z sieci neuronowych, a dane, które pobraliśmy z bazy danych nie zawierają zmiennej grupującej, umożliwiającej wybranie podzbioru uczącego i testowego, to konieczne jest opracowanie algorytmu umożliwiającego taki podział. Trudno sobie bowiem wyobrazić "ręczny" podział chociażby 5 000 przypadków w zbiorze danych.



Przeglądarka węzłów systemu *STATISTICA Data Miner*

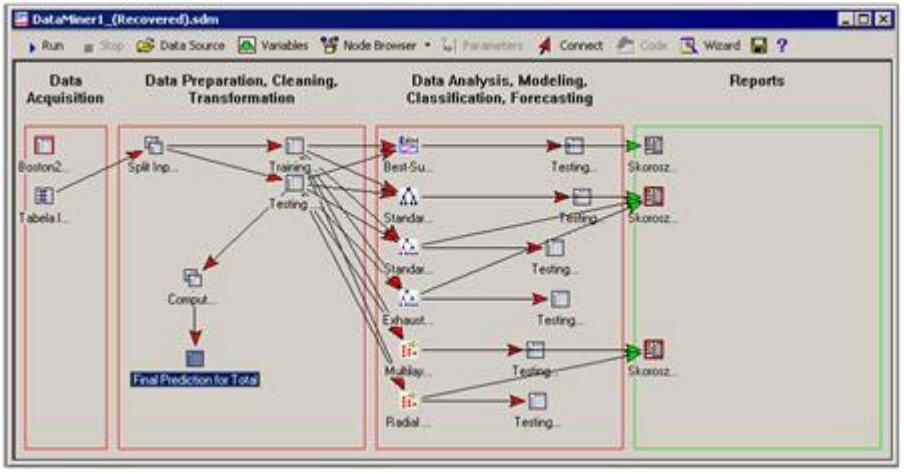
STATISTICA Data Miner umożliwia szeroki wybór algorytmów przetwarzania, czyszczenia i transformacji danych. Wszystkie te algorytmy łącznie z szerokim wyborem statystyk i wykresów dostępne są w przeglądarce węzłów. Korzystając z węzłów filtrowania danych możemy filtrować dane losowo lub względem określonej wartości minimalnej i maksymalnej, usuwać brakujące dane lub zastępować je średnią, eliminować i dobierać zmienne, które prawdopodobnie będą najlepszymi predyktorami dla bieżącego zbioru danych, a jednocześnie nie wprowadzą szumów do budowanych modeli. Prócz tego mamy do dyspozycji szereg algorytmów umożliwiających transponowanie, sortowanie,

rangowanie i standaryzowanie danych. Przy zagadnieniach regresyjnych i klasyfikacyjnych możemy korzystać z algorytmów podziału zbioru danych na próby uczącą i testową (np. dla sieci neuronowych) lub w przypadku wielu modeli zawartych w jednym projekcie, obliczać dla nich najlepsze predyktory. W każdym z powyżej opisanych przypadków jako wynik otrzymamy arkusz odpowiednio przetworzonych danych, który będzie wykorzystany w dalszych etapach projektu *data mining*.

W większości przypadków osoba korzystająca z modelu *data mining* połączonego z danymi, nie może modyfikować wartości danych znajdujących się w bazie danych. Wyobraźmy sobie prostą sytuację, że dane wprowadzane do bazy danych były przez dwie różne osoby. W kolumnie o nazwie "Płeć" jedna z osób wpisywała wartości tekstowe "Kobieta", "Mężczyzna", natomiast druga osoba kodowała płeć podając jedynie litery "K" i "M□. Wprowadzenie takich danych do modelu *data mining* może spowodować, że otrzymamy 4 a nie 2 klasy przypadków. Sytuacji takiej można jednak uniknąć stosując opisywane metody czyszczenia i transformacji danych. Algorytm polegałby wówczas na zamianie wartości "K" na "Kobieta" oraz "M" na "Mężczyzna". Wówczas dane w bazie danych pozostaną niezmienione, a dla potrzeb analiz zostanie utworzony poprawny zbiór danych lub, w przypadku przetwarzania danych po stronie serwera, będą one korygowane "w locie".

Analiza danych, modelowanie, klasyfikacja

Podobnie jak wszystkie programy z rodziny *STATISTICA*, również *STATISTICA Data Miner* ma otwartą architekturę obiektową (bazującą na modelu COM). Znajduje to odzwierciedlenie w budowanym projekcie *data mining*, w którym wszelkie operacje na danych obrazowane są przez węzły analiz czy wykresów (obiekty). Tworząc modele *data mining*, przenosimy poszczególne węzły z przeglądarki obiektów do przestrzeni roboczej programu. Po połączeniu ich otrzymujemy kompletny projekt, który można zapisać na dysku komputera i przesłać współpracownikom lub klientom.



Przykładowy projekt *data mining* w środowisku *STATISTICA Data Miner*

Dzięki otwartej architekturze system można bardzo łatwo wbudować w istniejąca strukturę informatyczną. *STATISTICA Data Miner* może np. pracując w tle generować prognozy i umieszczać je w bazie danych, w systemie CRM lub w Intranecie organizacji.

Kluczowym elementem każdego projektu *data mining* jest wybór analiz odpowiednich dla typu i treści danych oraz celów projektu. Techniki *data mining* możemy podzielić na następujące grupy:

* opis danych,
* klasyfikacja,
* modelowanie zmiennej ciągłej,
* prognozowanie,
* analiza koszykowa (asocjacji),
* analiza skupień (segmentacja.

Wszystkie procedury analityczne umożliwiające realizację zadań *data mining* dostępne są w systemie *STATISTICA Data Miner*. Ponadto użytkownicy o niekonwencjonalnych wymaganiach analitycznych mogą przy pomocy wbudowanego w system standardowego języka *STATISTICA* Visual Basic, tworzyć od podstaw lub modyfikować wszystkie węzły przetwarzania i analizy danych. Do systemu można też dodawać własne procedury napisane np. w jeżyku C++ lub Java.

System *STATISTICA Data Miner* dysponuje wszystkimi procedurami analitycznymi i graficznymi programu *STATISTICA*. Ponadto w skład systemu wchodzą najnowsze, wyrafinowane techniki analizy danych: *Uogólnione modele addytywne*, *Ogólne modele CHAID*, *Drzewa klasyfikacyjne i regresyjne*, *Drzewa interakcyjne (C&RT, CHAID)*, *Sieci neuronowe*, *Interakcyjne drążenie danych*, *Obliczanie dobroci dopasowania*, *Dobór i eliminacja zmiennych*, *Analiza koszykowa, Algorytm rekurencyjnego podziału przestrzeni cech do budowy modelu regresyjnego w postaci krzywych składanych (MAR Splines)*.

Użytkownicy systemu mogą skorzystać z kreatora *STATISTICA Data Miner*, który prowadzi użytkownika przez proces określania analiz począwszy od zdefiniowania źródła danych, przez wybór zmiennych, analiz po raporty i wizualizacje zdobytych w procesie *data mining* informacji. Ponadto w systemie dostępne są gotowe projekty dla poszczególnych typów zadań, do który wystarczy podłączyć źródło danych, aby móc z nich skorzystać.

Po zbudowaniu modelu *data mining*, kiedy są określone wszystkie jego parametry, możemy przenieść lub podłączyć jego kod do własnego programu (napisanego w Visual Basic, C/C++, Java, czy dowolnym języku obiektowym) i w ten sposób korzystać z gotowego rozwiązania *data mining*. Uzyskiwane modele możemy zapisywać również w jeżyku XML, zgodnie ze standardem PMML.

Często w poszukiwaniu dobrego modelu *data mining* stosujemy wiele różnych metod analitycznych. Np. w zagadnieniach klasyfikacyjnych porównujemy działanie sieci

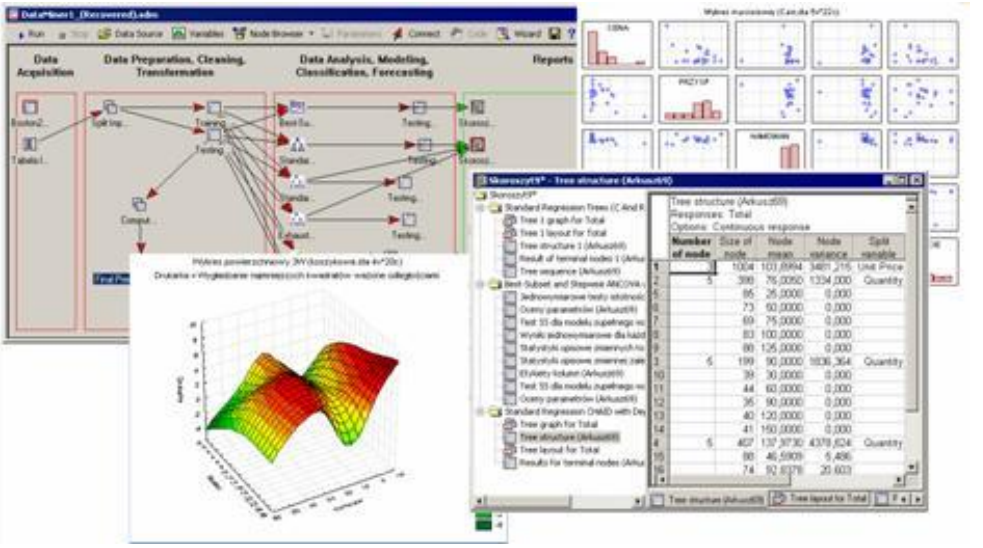
neuronowych, drzew klasyfikacyjnych lub ogólnych modeli analizy dyskryminacji, otrzymując podobne, ale jednak nieco różne wyniki. Który z modeli jest najlepszy? A może wszystkie klasyfikują jednakowo dobrze? Takie pytania nurtują nie tylko początkujących analityków, ale również zaawansowanych "data minerów". Odpowiedź na powyższe pytania znajduje się w module *STATISTICA Data Miner - Obliczanie dobroci dopasowania*. Moduł ten pozwala na obliczanie różnych statystyk określających dobroć dopasowania dla odpowiedzi o charakterze ciągłym i skategoryzowanym (w przypadku zagadnień regresyjnych i klasyfikacyjnych).

Wizualizacja i rozpowszechnianie wyników

Szeroki wybór technik i analiz *data mining* wymaga równie szerokich możliwości graficznego przedstawienia wyników ich działania. Czasami do interpretacji wyników, czy opisania zależności zmiennych wystarczy prosty histogram lub wykres rozrzutu, ale zazwyczaj procesy *data mining* dotyczą bardzo złożonych i zróżnicowanych zagadnień. Konieczna jest więc w niektórych sytuacjach graficzna eksploracja danych.

STATISTICA Data Miner zawiera wszechstronny zestaw narzędzi przeznaczonych do graficznej eksploracji i analizy danych oraz służących do identyfikacji relacji, trendów i błędów systematycznych "ukrytych" w nieuporządkowanych zbiorach danych. Techniki graficznego *data mining* obejmują dopasowywanie i wykreślanie funkcji, wygładzanie danych, nakładanie i łączenie wielu obrazów, interakcyjne rozdzielanie danych na kategorie, podział i łączenie podzbiorów danych na wykresach, agregowanie danych, identyfikacje i zaznaczanie podzbiorów danych spełniających określone warunki, wykreślanie przedziałów i obszarów (elips) ufności, generowanie wykresów mozaikowych, płaszczyzn spektralnych, rzutowanych warstwic, techniki redukcji obrazowanych danych, interakcyjne (i płynne) obracanie wykresów trójwymiarowych, selektywne podświetlanie określonych serii i bloków danych, interakcyjne wybieranie trójwymiarowych obszarów na wykresie, analityczne narzędzie powiększania umożliwiające interakcyjny wybór fragmentu wykresu i przedstawienie go na oddzielnym wykresie oraz wiele innych.

Tak szeroki wybór graficznych technik prezentacji wyników i graficznego *data mining* umożliwia dostosowanie wizualizacji danych do indywidualnych potrzeb i oczekiwań użytkownika.



Projekt *data mining* ze skoroszytem zawierającym wykresy i tabele podsumowujące

Wyniki analiz możemy kierować do skoroszytu, oddzielnych węzłów (każda tabela lub wykres będzie stanowił odrębny obiekt w przestrzeni roboczej projektu), lub bezpośrednio do raportu. Tak sformatowane wyniki projektu *data mining*, lub ich część, możemy zapisać w postaci elektronicznej i przesłać innym współpracownikom lub klientom pocztą elektroniczną. Alternatywnym i bardziej globalnym rozwiązaniem jest umieszczenie ich w Internecie lub lokalnym Intranecie w postaci plików html.

Możliwość prezentacji samych wyników w środowisku Internetowym to nie wszystko. *STATISTICA Data Miner* dzięki opcjonalnej aplikacji *STATISTICA Enterprise Server* (*www.statsoft.pl/webserver.html*) umożliwia wykonywanie wszystkich operacji *data mining*, w oknie przeglądarki internetowej na dowolnym komputerze połączonym z Internetem. Dzięki temu można pracować nad projektami przez Internet i współpracować zarówno z osobami w tym samym biurze, jak i na innym kontynencie. Taka konfiguracja sprawia, że użytkownik nie musi mieć na swoim komputerze zainstalowanego systemu *STATISTICA Data Miner*, natomiast korzysta ze wszystkich jego funkcji przez przeglądarkę internetową.

Przykład rozwiązania zagadnienia predykcyjnego za pomocą technik *Data mining*

W przykładzie prezentowane jest zastosowanie technik *data mining* do zagadnienia predykcyjnego. Do wykonania analizy zostanie wykorzystany program *STATISTICA Data Miner*.

Opis problemu i przykładowych danych

Celem przeprowadzanych analiz jest budowa modeli wyjaśniających wpływ różnych predyktorów na ceny nieruchomości w pewnym rejonie Kalifornii (por. [6]). Ceny nieruchomości są wyrażone poprzez medianę wartości nieruchomości (**Mediana wartości**

domów; pełni ona w analizie rolę zmiennej zależnej). Wśród dostępnych potencjalnych predyktorów występują następujące zmienne:

* **Długość geograficzna** i **Szerokość geograficzna**; określające położenie rejonu w którym znajdują się nieruchomości
* **Mediana wieku domu**; "wiek" nieruchomości, wyrażony w latach
* **Pokoje ogółem**, łączna liczba pokoi w branych pod uwagę nieruchomościach
* **Sypialnie ogółem**, łączna liczba sypialni w branych pod uwagę nieruchomościach
* **Liczba osób**, łączna liczba osób zamieszkujących nieruchomości
* **Liczba mieszkań**, łączna liczba mieszkań w branych pod uwagę nieruchomościach
* **Mediana dochodu**, wartość mediany dochodu osób zamieszkujących nieruchomości wyrażona w tys. dolarów.

Oprócz budowy modeli drugim celem niniejszego przykładu jest praktyczna prezentacja sposobu przeprowadzania bardziej złożonych analiz z użyciem narzędzi analitycznych zawartych w programie ***STATISTICA Data Miner*** [5].

Przygotowanie danych do analiz

Dane do opisywanego przykładu są zapisane w pliku *Kalifornia.sta*. Rozpoczynając budowanie projektu *data mining*, w panelu *Źródło danych* wskazujemy plik danych o nazwie *Kalifornia.sta*. W tym celu możemy skorzystać z przycisku *Data Source*. Następnie, w oknie *Select dependent variables and predictors*, które pojawi się na ekranie, wskazujemy zmienną zależną i predyktory (zmienne objaśniające). W naszym przykładzie, w charakterze zmiennej zależnej ciągłej użyjemy zmiennej o nazwie **Mediana wartości domu** natomiast jako predyktory ciągłe wskazujemy zmienne: **Długość geograficzna**, **Szerokość geograficzna**, **Mediana wieku domu**, **Pokoje ogółem**, **Sypialnie ogółem**,**Liczba osób**, **Liczba mieszkań** oraz **Mediana dochodu**. Kliknięciem przycisku *OK* akceptujemy dokonane wybory a następnie w oknie *Select dependent variables and predictors* jeszcze raz klikamy przycisk *OK*. Spowoduje to powrót do obszaru roboczego projektu.

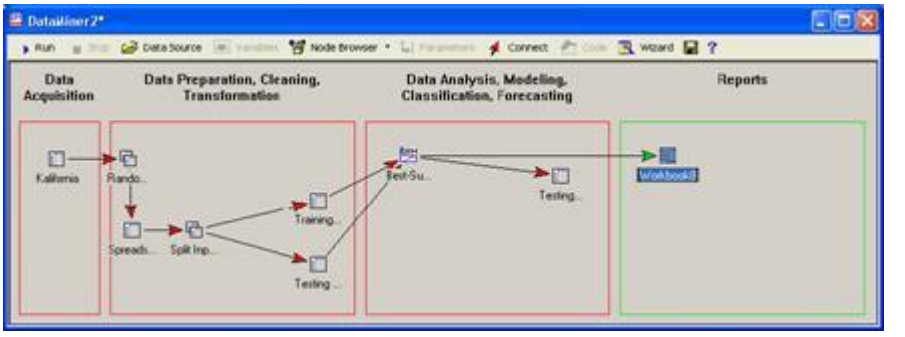
Ze względu na czas obliczeń wyjściowy plik danych ograniczymy do 10% przypadków. W tym celu korzystamy z przycisku *Node Browser* i w lewym panelu okna, które pojawi się na ekranie rozwijamy katalog *Data Cleaning and Filtering*, a następnie w prawym panelu klikamy dwukrotnie lewym przyciskiem myszy węzeł analizy o nazwie *Random Sample Filtering*. Za pomocą tego węzła możemy wybrać losowy podzbiór przypadków o żądanej liczebności. Powracamy do obszaru projektu, klikamy prawym przyciskiem myszy ikonkę oznaczającą wstawiony węzeł oraz w podręcznym menu wybieramy opcję *Edit parameters*. W oknie, które pojawi się na ekranie klikamy kartę *General* i w polu *Percent of Cases* wprowadzamy wartość *10*. Kliknięciem przycisku OK akceptujemy dokonany wybór i wracamy ponownie do obszaru roboczego projektu. Aby wykonać żądane zadanie używamy przycisku paska narzędzi*Run*. Program wykonuje odpowiednie obliczenia i

umieszcza wyniki w nowym arkuszu. Możemy je przejrzeć klikając prawym przyciskiem ikonkę nowego arkusza oraz wybierając w podręcznym menu opcję *View Document*.

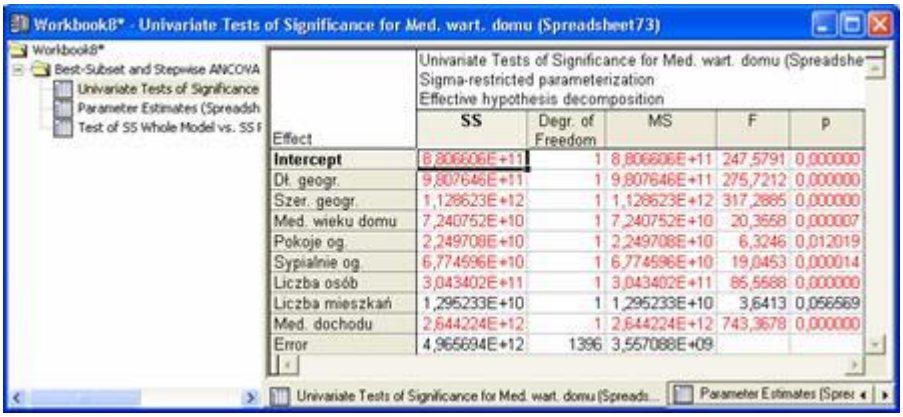
Kolejną czynnością, którą przeprowadzimy będzie podział uzyskanego podzbioru danych na dwie części: próbę uczącą (training sample) i próbę testową (testing sample). Próba ucząca będzie wykorzystywana do szacowania parametrów tworzonych modeli natomiast próba testowa będzie wykorzystywana do testowania uzyskanych rozwiązań. Do próby testowej weźmiemy 25% przypadków (spośród wybranych wcześniej 10%). Odpowiedni węzeł analizy; *Split data into Training and Testing Samples* przywołujemy z katalogu *Regression Modeling and Multivariate Exploration* w *Przeglądarce węzłów*. Po powrocie do obszaru roboczego projektu klikamy prawym przyciskiem myszy ikonkę oznaczającą wstawiony węzeł oraz w podręcznym menu wybieramy opcję *Edit parameters*. W oknie, które pojawi się na ekranie klikamy kartę *General* i w polu *Approximate percent of cases for testing:* wprowadzamy wartość *25*.

Budowa modeli predyktywnych

Po etapie wstępnego przygotowania danych przystąpimy do zasadniczej części analizy. Przy budowie pierwszego modelu wykorzystamy regresję wieloraką. W tym celu do próby uczącej i testowej podpinamy węzeł analizy o nazwie *Best-Subset and Stepwise ANCOVA with Deployment*, który znajduje się w katalogu *Regression Modeling and Multivariate Exploration*. Klikając przycisk *Run* umieszczony na pasku narzędzi uruchamiamy odpowiednią analizę.

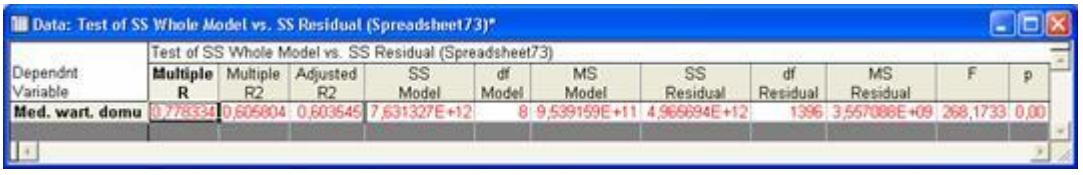


Aby obejrzeć otrzymane wyniki klikamy dwukrotnie ikonkę symbolizującą skoroszyt z wynikami analizy. Na ekranie pojawi się okno skoroszytu, pokazane na poniższym zrzucie. Zaznaczając odpowiednią pozycję w lewym panelu skoroszytu możemy obejrzeć odpowiednie wyniki.



Effect	SS	Degr. of Freedom	MS	F	p
Intercept	8,806606E+11	1	8,806606E+11	247,5791	0,000000
Dł. geogr.	9,807646E+11	1	9,807646E+11	275,7212	0,000000
Szer. geogr.	1,128623E+12	1	1,128623E+12	317,2885	0,000000
Med. wieku domu	7,240752E+10	1	7,240752E+10	20,3558	0,000007
Pokoje og.	2,249708E+10	1	2,249708E+10	6,3246	0,012019
Sypialnie og.	6,774596E+10	1	6,774596E+10	19,0453	0,000014
Liczba osób	3,043402E+11	1	3,043402E+11	85,5588	0,000000
Liczba mieszkań	1,295233E+10	1	1,295233E+10	3,6413	0,056569
Med. dochodu	2,644224E+12	1	2,644224E+12	743,3678	0,000000
Error	4,965694E+12	1396	3,557088E+09		

I tak wyniki zamieszczone w arkuszu *Univariate Tests of Significance* sugerują, że wszystkie dostępne w pliku danych predyktory powinny zostać uwzględnione w modelu. Poniżej przedstawiono okno z wybranymi miarami dobroci dopasowania modelu. Wynika z nich, że zbudowany model wyjaśnia nieco ponad 60% oryginalnej zmienności zmiennej zależnej.

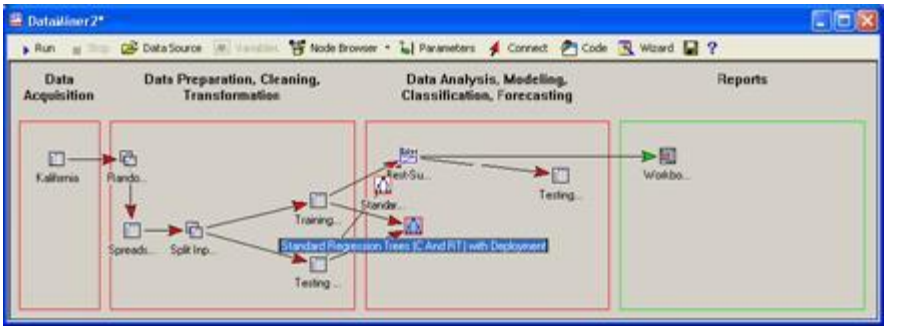


Dependnt Variable	Multiple R	Multiple R2	Adjusted R2	SS Model	df Model	MS Model	SS Residual	df Residual	MS Residual	F	p
Med. wart. domu	0,778534	0,605804	0,603545	7,631327E+12	8	9,539159E+11	4,965694E+12	1396	3,557088E+09	268,1733	0,00

Kolejne okno zawiera oceny parametrów strukturalnych modelu, które są wykorzystywane przy tworzeniu prognozy. Są w nim również podawane standaryzowane oceny współczynników regresji dla każdego z predyktorów. Z ich wartości wynika, że stosunkowo największy wpływ na wartości zmiennej zależnej wykazują zmiany długości i szerokości geograficznej rejonu, w którym były ulokowane nieruchomości.

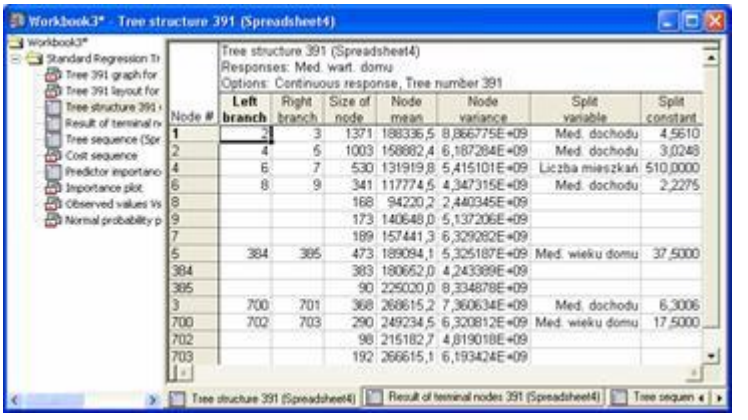
Effect	Med. wart. domu Param.	Med. wart. domu Std.Err	Med. wart. domu t	Med. wart. domu p	-95,00% Cnf.Lmt	+95,00% Cnf.Lmt	Med. wart. domu Beta (ß)
Intercept	-3279295	208444,2	-15,7346	0,000000	-3688893	-2870098	
Dł. geogr.	-39375	2371,3	-16,6049	0,000000	-44026	-34723	-0,811032
Szer. geogr.	-39479	2216,4	-17,8126	0,000000	-43827	-35132	-0,880706
Med. wieku domu	758	167,9	4,5117	0,000007	429	1087	0,088448
Pokoje og.	-7	2,8	-2,5149	0,012019	-12	-2	-0,157283
Sypialnie og.	101	23,1	4,3641	0,000014	56	146	0,447987
Liczba osób	-37	4,0	-9,2498	0,000000	-45	-29	-0,430272
Liczba mieszkań	47	24,7	1,9082	0,056569	-1	96	0,188258
Med. dochodu	38556	1414,1	27,2648	0,000000	35782	41330	0,629790

Kolejny model zostanie przygotowany za pomocą modułu *Standard Regression Trees (C and RT) with Deployment*. Zostanie wykorzystana procedura drzew regresyjnych. Jest ona również dostępna w katalogu *Regression Modeling and Multivariate Exploration*. Tak jak poprzednio aby zastosować ten model do naszych danych "podpinamy" w obszarze roboczym projektu *data mining* zbiór uczący i testowy do ikony oznaczającej żądaną analizę. Projekt analizy wygląda teraz tak jak na poniżej zamieszczonej ilustracji.



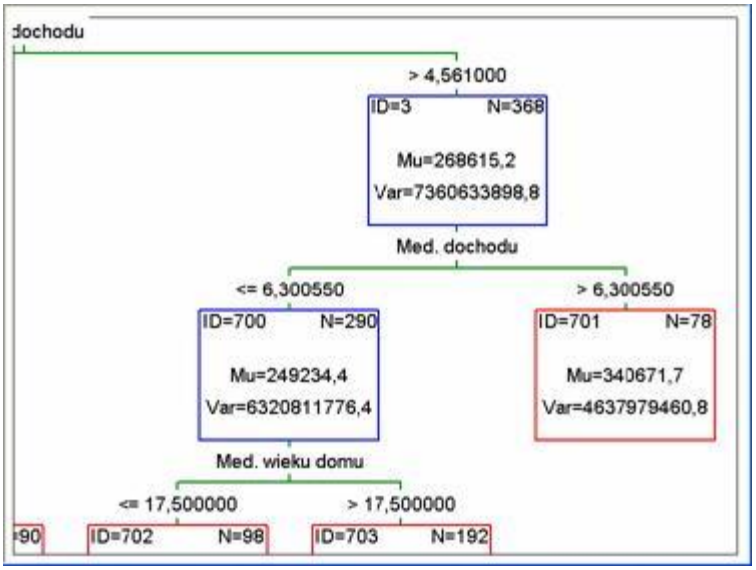
Przed wykonaniem odpowiednich obliczeń zmienimy jeszcze zakres uzyskiwanych wyników. W tym celu klikamy prawym przyciskiem myszy ikonę oznaczającą wybraną przez nas analizę i wybieramy opcję *Edit parameters*. Na karcie *General* tego okna w polu *Detail of computed results reported* wybieramy pozycję *Comprehensive*. Oprócz tego na karcie *V-Fold Crossvalidation* zaznaczamy opcję v-krotnej oceny krzyżowej (pozostawiając domyślne parametry). Tak jak poprzednio aby wykonać odpowiednie obliczenia dotyczące tylko nowo wstawionego węzła klikamy prawym przyciskiem myszy w obrębie obszaru roboczego projektu i w podręcznym menu wybieramy opcję*Run dirty nodes*.

Podobnie jak przy poprzedniej analizie, aby obejrzeć otrzymane wyniki klikamy dwukrotnie ikonkę symbolizującą skoroszyt z wynikami analizy. Na ekranie pojawi się okno skoroszytu, pokazane na poniższym zrzucie. Zaznaczając odpowiednią pozycję w lewym panelu skoroszytu możemy obejrzeć żądane wyniki.

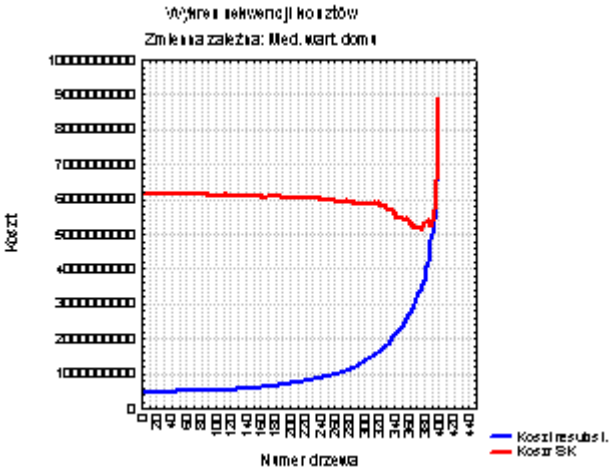


Node #	Left branch	Right branch	Size of node	Node mean	Node variance	Split variable	Split constant
1	2	3	1371	188336,5	8,866775E+09	Med. dochodu	4,5610
2	4	5	1003	158882,4	6,187284E+09	Med. dochodu	3,0248
4	6	7	530	131919,8	5,415101E+09	Liczba mieszkań	510,0000
6	8	9	341	117774,5	4,347315E+09	Med. dochodu	2,2275
8			168	94220,2	2,440345E+09		
9			173	140648,0	5,137206E+09		
7			189	157441,3	6,329282E+09		
5	384	385	473	189094,1	5,325187E+09	Med. wieku domu	37,5000
384			383	180652,0	4,243389E+09		
385			90	225020,0	8,334878E+09		
3	700	701	368	268615,2	7,360634E+09	Med. dochodu	6,3006
700	702	703	290	249234,5	6,320812E+09	Med. wieku domu	17,5000
702			98	215182,7	4,819018E+09		
703			192	266615,1	6,193424E+09		

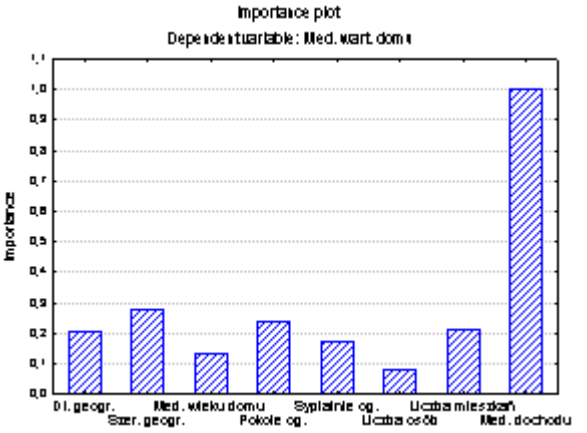
Przeglądanie wyników rozpoczniemy od drzewa regresyjnego. Przedstawia ono reguły decyzyjne występujące przy podziale przypadków. Fragment uzyskanego drzewa przedstawia poniższy zrzut.



Jako kolejny wynik analizy obejrzymy wykres liniowy sekwencji kosztów. Na wykresie tym znajduje się koszt sprawdzianu krzyżowego oraz koszt resubstytucji dla każdego przyciętego drzewa. Punkt przecięcia wykreślonych linii wskazuje optymalne drzewo. W naszym przykładzie jest to drzewo o numerze 408, które posiada 11 węzłów końcowych. Na zamieszczonym poniżej wykresie przedstawiono wykres sekwencji kosztów.

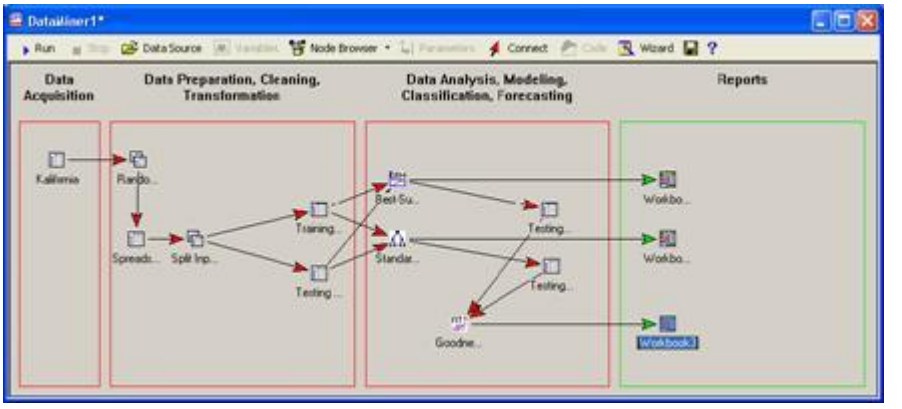


Możemy także obejrzeć wykres przedstawiający wielkość względnego wpływu poszczególnych zmiennych. Jak widać z tego punktu widzenia najważniejszym predyktorem okazuje się zmienna **Mediana dochodu.** Wykres ten został pokazany poniżej.

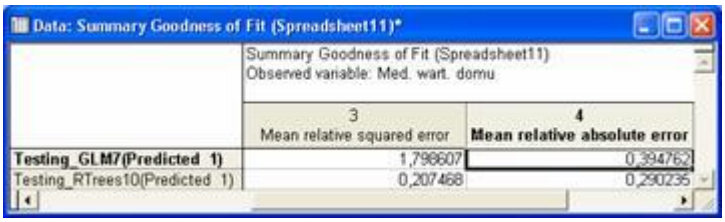


Ocena uzyskanych rozwiązań

W kolejnej części analizy ocenimy dobroć dopasowania uzyskanych wcześniej modeli. W tym celu wykorzystamy węzeł o nazwie *Goodness of Fit for Multiple Inputs*, który znajduje się w katalogu *Statistics/Data Mining/Goodness of Fit*. Z węzłem tym połączymy kolejno arkusze, zawierające wartości obserwowane i prognozowane dla każdego ze stosowanych modeli. Przed połączeniem musimy wskazać dla każdego arkusza wybór zmiennych. Jako zmienną zależną wskazujemy wartość obserwowaną a jako zmienną niezależną wartość prognozowaną. Możemy także określić zakres uzyskiwanych wyników. Możemy go ustalić po kliknięciu prawym przyciskiem myszy na ikonie symbolizującej odpowiedni węzeł i wybraniu opcji *Edit parameters* oraz karty *Continuous*. Na karcie tej zaznaczamy wszystkie dostępne przycisku opcji. Nasz projekt analizy wygląda teraz tak jak na poniższym zrzucie.

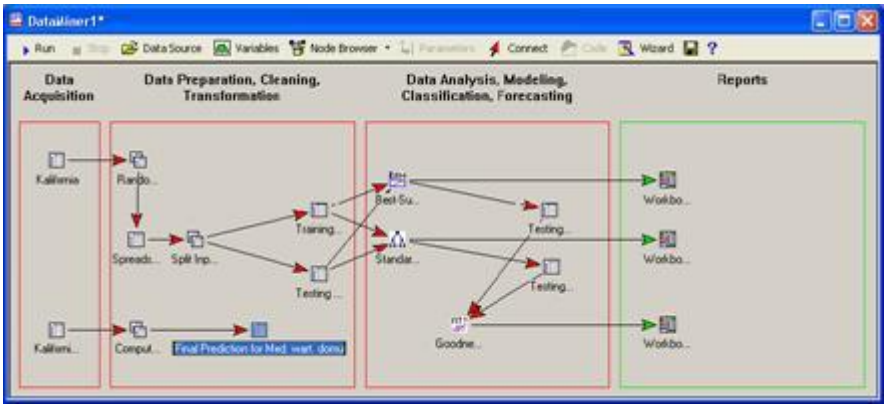


- Teraz możemy już uruchomić odpowiednie obliczenia. Wszystkie wyniki obliczeń zawarte są w jednym arkuszu. Jego fragment przedstawiamy poniżej. Biorąc pod uwagę wartość średniego względnego błędu kwadratowego lub wartość średniego względnego błędu absolutnego możemy stwierdzić, że stosunkowo najlepsze prognozy uzyskujemy w przypadku zastosowania sieci neuronowej.

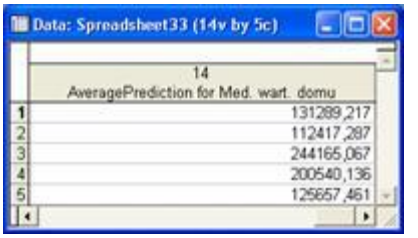


Zastosowanie uzyskanych modeli do prognozowania wartości nowych przypadków

Na koniec analizy zbudujemy prognozę w oparciu o wyniki wszystkich zastosowanych modeli. W tym celu wykorzystamy węzeł o nazwie *Compute Best Prediction from All Modules*, który znajduje się w katalogu *Regression Modeling and Multivariate Exploration*. Prognozy zbudujemy w oparciu o dane zawarte w pliku *Kalifornia_pred*. Plik ten zawiera tylko dane dla wszystkich występujących w naszej analizie predyktorów. Najpierw musimy wstawić ten plik do obszaru roboczego projektu. Następnie łączymy go strzałką z węzłem *Compute Best Prediction from All Modules*. Przed wykonaniem odpowiednich obliczeń klikamy ikonkę wstawionego pliku prawym przyciskiem myszy i wybieramy opcję *Variable selection*. Wybór zmiennych pozostaje taki sam jak poprzednio. Musimy tylko pamiętać aby koniecznie zaznaczyć pole wyboru *Data for deployment project; do not re-estimate model* umieszczonej w dolnej części okna. Analizę uruchamiamy za pomocą opcji *Run dirty nodes*. Ostateczny wygląd naszego projektu analizy przedstawia poniższy zrzut.



Klikając prawym przyciskiem ikonkę *Final Prediction for Med. wart. domu* i wybierając opcję *View document* możemy obejrzeć średnią wartość prognozy wyliczoną w oparciu o wyniki uzyskane przez wszystkie trzy zastosowane modele. Wartości te dla wybranych pięciu przypadków przedstawia poniższy zrzut.



Literatura

1. *Marketing i statystyka*, StatSoft Polska, Kraków, 1999 (dostępny również w postaci elektronicznej na stronie *www.statsfot.pl/czytelnia.html*)

2. Berry M. J. A., Linoff G., 1997, *Data mining* Techniques, for Marketing, Sales and Customer Support, John Wiley & Sons, Inc.

3. Han J., Kamber M., 2001, *Data mining*, Concepts and Techniques, Morgan Kaufmann Publishers

4. Gatnar E., 2001, Nieparametryczna metoda dyskryminacji i regresji, PWN Warszawa.

5. *STATISTICA Data Miner* Manual, StatSoft, Inc., 2002.

6. *Hastie T., Tibshirani R., Friedman J.*, 2002, The Elements of Statistical Learning, Springer

