



CZY STOSOWANIE METOD DATA MINING MOŻE PRZYNOŚIĆ KORZYŚCI W BADANIACH NAUKOWYCH?

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

Obecny wiek charakteryzuje się pojawianiem się coraz większej liczby skomplikowanych problemów z pogranicza wielu dziedzin nauki. Co więcej gromadzone są ogromne ilości danych na temat obserwowanych zjawisk, z których ze względu na ich wolumen coraz trudniej jest wydobyć wiedzę. Ilość danych, które należy przeanalizować, często powoduje, że techniki oferowane przez tradycyjną statystykę okazują się niewystarczające do należytego ich opracowania. Badacze, chcąc wydobyć wiedzę zawartą w danych, coraz częściej sięgają więc po narzędzia służące do eksploracji danych (data mining).

Data mining nie musi być oczywiście wykorzystywany do bardzo dużych zbiorów danych, jego przydatność w badaniach naukowych może ujawnić się także w przypadku trudności ze skonstruowaniem teoretycznego modelu badanego zjawiska. W takiej sytuacji modele data mining, które oparte są jedynie na danych, mogą umożliwić badaczowi wyciągnięcie istotnych merytorycznie wniosków na temat analizowanego problemu.

Metody data mining zazwyczaj nie wymagają spełnienia żadnych formalnych założeń co do rozkładu, wariancji czy innych charakterystyk badanych zmiennych, dlatego też mogą zostać użyte w zasadzie dla dowolnych danych, choć oczywiście podobnie jak w przypadku statystyki modele dobrej jakości wymagają, aby dane przed analizą zostały odpowiednio przygotowane i oczyszczone.

Bardzo ciekawe i inspirujące pole wykorzystania technik data mining zostało zaprezentowane w artykule profesora Tadeusiewicza [5], który zaproponował wykorzystanie technik data mining w celu odkrycia nieznanymi wcześniej związków i zależności zawartych w „wyeksploatowanych” danych empirycznych, na podstawie których badacze wyciągnęli już wcześniej wnioski.

Koncepcja data mining

Data mining jest relatywnie młodą interdyscyplinarną dziedziną powstałą w efekcie połączenia wiedzy z zakresu technik baz danych, statystyki, sztucznej inteligencji oraz nauk społeczno-ekonomicznych. Większość metod data mining swoje źródło ma właśnie



w badaniach nad sztuczną inteligencją bądź powstało w wyniku rozwijania wielowymiarowych metod statystycznych.

Powstanie tego typu metod jest silnie związane z rozwojem technik komputerowych – znaczenie metod statystycznych wymagających zaawansowanych obliczeń numerycznych zaczęło rosnąć w drugiej połowie lat osiemdziesiątych wraz z pojawieniem się coraz powszechniejszej możliwości wykorzystania komputerów do obliczeń. Dzięki nim stało się możliwe wykorzystanie metod statystycznych do rozwiązywania skomplikowanych, wielowymiarowych problemów.

Kolejnym czynnikiem było zainteresowanie statystyków metodami zaliczanymi do metod uczenia maszynowego (spowodowane m.in. rozwojem sieci neuronowych oraz dopracowaniem metod opartych na drzewach decyzyjnych). Metody te przestały być również wykorzystywane jedynie na polu obliczeń numerycznych i sztucznej inteligencji. W szczególności zaczęły być używane w marketingu opartym na bazach danych, a następnie swoje zastosowania znalazły w praktycznie każdej dziedzinie biznesu czy nauki, w której potrzebna jest analiza danych.

W początkowym etapie rozwoju technik data mining oczekiwano, że techniki te w pełni zautomatyzują pracę analityka, że wystarczy tak naprawdę jedno kliknięcie myszką i możliwe będzie uzyskanie interesujących wyników, wyszukanych spośród milionów bajtów danych. Praktyka kolejnych lat mocno zweryfikowała jednak ten pogląd. Okazało się, że uzyskanie wartościowych rezultatów jest zajęciem bardzo czasochłonnym i pracochłonnym wymagającym od uczestników projektu sporej wiedzy i zaangażowania.

Ewolucję tę można zaobserwować na przykładzie definicji data mining proponowanych przez M. A. Berry'ego oraz G. Linoffa propagatorów data mining i autorów książek o tej tematyce. Otóż w swojej książce dotyczącej data mining [1] wydanej w 1997 r. zaproponowali oni następującą definicję: „*Data mining jest procesem badania i analizy dużych ilości danych metodami automatycznymi lub półautomatycznymi w celu odkrycia znaczących wzorców i reguł*”. Jednak już w swojej kolejnej o trzy lata późniejszej książce [2] zaproponowali zmianę tej definicji i usunięcie frazy „*metodami automatycznymi lub półautomatycznymi*”. Podobnie w wielu innych publikacjach z ostatnich lat często pojawia się ostrzeżenie, aby nie traktować data mining jak „magicznej” metody, która sama, bez żadnego udziału człowieka jest w stanie wydobyć z danych użyteczną wiedzę.

Data mining a statystyka

Celem zarówno statystyki, jak i data mining jest analiza danych. Jednak podejście do analizy danych stosowane w data mining różni się od tego stosowanego w statystyce. Mówiąc najogólniej, celem statystyki jest zwykle weryfikacja pewnych modeli opartych na teoretycznej wiedzy (znany jest ogólny charakter zależności, a my staramy się oszacować parametry) lub weryfikacja pewnych hipotez badawczych. Data mining stosujemy natomiast wtedy, gdy nie znamy ogólnego kształtu zależności, które pragniemy modelować, chcemy odnaleźć pewne nieznane interesujące związki. Dodatkowo w data mining główny nacisk

kładzie się na praktyczne zastosowania, często zaniedbując mechanizmy odpowiedzialne za modelowane zjawiska lub procesy (w data mining bardziej niż poziom istotności interesuje nas skuteczność modelu). Dlatego też w data mining dopuszcza się wykorzystanie metod działających na zasadzie czarnej skrzynki (np. sieci neuronowych).

Drugą różnicą pomiędzy statystyką a data mining jest pochodzenie danych oraz sposób podejścia do procesu badawczego. W przypadku statystyki po zdefiniowaniu problemu badawczego, najpierw konstruujemy teoretyczny model zjawiska, a następnie planujemy eksperyment i gromadzimy na jego podstawie dane - głównym celem gromadzenia danych jest analiza statystyczna. W wyniku analizy danych następuje weryfikacja sformułowanych we wcześniejszym etapie hipotez.



Rys. 1. Tradycyjny proces badawczy.

W przypadku procesu data mining, podobnie jak w poprzednim przypadku, zaczynamy od definiowania problemu badawczego. Natomiast kolejnym krokiem jest już analiza danych, którymi zwykle dysponujemy już na wstępie. Są to na przykład dane gromadzone na potrzeby sprawozdawczości bądź bieżącej działalności danej instytucji lub też dane pochodzące z eksperymentu dla innego problemu badawczego. Zbudowany model jest więc oparty nie na naszej wiedzy, lecz jedynie na danych. Po jego zbudowaniu oceniamy i interpretujemy uzyskane wyniki.

Używane do analiz dane różnią się zwykle (choć nie jest to wymóg) wolumenem. W przypadku data mining mogą to być nawet ekstremalnie duże zbiory danych.



Rys. 2. Proces data mining.

Statystycy przez długi czas określali data mining jako łowienie w danych lub kopanie w danych. Określenia te były oczywiście używane w negatywnym kontekście. Krytyka ta była spowodowana dwoma kwestiami. Pierwszy zarzut dotyczył strategii budowy modeli, polegającej na opracowaniu szeregu konkurujących ze sobą modeli i wyborze najlepszego. Krytycy takiego podejścia twierdzili, że zawsze można opracować model, chociaż skomplikowany, który dobrze dopasuje się do danych. Drugi zarzut mówił, że w olbrzymiej ilości danych bardzo łatwo można odnaleźć pozorne zależności, które w rzeczywistości nie występują [4] (w przypadku podejścia statystycznego duża liczba obserwacji może spowodować, że nawet stosunkowo słabe efekty mogą się okazać statystycznie istotne).

Chociaż te krytyczne uwagi są warte rozważenia, to jednak warto zwrócić uwagę, że nowoczesne metody data mining kładą ogromny nacisk na możliwości generalizacji zbudowanych modeli. Oznacza to, że podczas wyboru najlepszego modelu bierze się pod uwagę zarówno jego zdolności do aproksymacji, jak i do generalizacji, a także jego stopień skomplikowania. Jeśli chodzi o drugi zarzut, to nie można zignorować faktu, że wiele ważnych informacji nie jest znanych przed analizą i nie może być użytych w formułowaniu hipotezy badawczej.

Dodatkowo praktyka ostatnich lat pokazuje, że szereg problemów badawczych, których statystyka nie potrafiła rozwiązać (np. ze względu na wolumen danych bądź wymogi formalne) znalazło swoje rozwiązanie za pomocą metod data mining.

Pomimo istotnych różnic w obu podejściach warto pamiętać, że granice pomiędzy nimi są bardzo nieostre. Dodatkowo z jednej strony bardzo mocno rozwijają się metody statystyczne oparte na skomplikowanych metodach numerycznych, z drugiej strony podejmowane są próby „wtłoczenia” metod data mining w ramy klasycznej statystyki.



Zadania data mining

Zadania, jakie mogą być rozwiązywane za pomocą metod data mining, niemal w całości pokrywają się z analogicznie zdefiniowanymi problemami rozwiązywanymi przy pomocy metod statystycznych.

Popularnym typem zadania data mining jest segmentacja. Segmentacja polega na podziale niejednorodnej grupy obiektów (np. klientów) na grupy. Wszystkie elementy znajdujące się w tej samej grupie uważane są za podobne do siebie, elementy znajdujące się w różnych grupach są różne. Kluczowym elementem analizy jest określenie optymalnej liczby segmentów. Liczba segmentów zależy od tego, jak zróżnicowane są analizowane przez nas obiekty, oraz od przyjętych przez nas kryteriów praktycznych (np. przyjmujemy, że segmentów nie może być więcej niż 6).

Zadanie to może zostać zrealizowane przez szereg metod data mining, takich jak: sieci neuronowe Kohonena, metodę k-średnich czy EM. Niemniej jednak to samo zadanie można zrealizować za pomocą tradycyjnej metody aglomeracyjnej (grupowanie metodą Warda), choć istotnym ograniczeniem jej stosowania jest wolumen danych – jeśli nasze dane zawierają więcej niż 200-300 przypadków, dendrogram aglomeracji staje się nieczytelny.

Drugim popularnym rodzajem zadania data mining jest budowa modeli predykcyjnych. Same modele tego typu możemy podzielić na kilka grup w zależności od rodzaju zmiennej zależnej (objaśnianej). Jeśli zmienna zależna jest zmienną ilościową, wtedy mówimy o modelowaniu problemu regresyjnego. Jeśli zmienna zależna jest jakościowa, mamy do czynienia z klasyfikacją. Szczególnym przypadkiem regresji, gdzie pomiary pewnej cechy są wykonywane w kolejnych momentach czasu, są szeregi czasowe. Niezależnie od rodzaju modelu predykcyjnego, zadanie tego typu zawsze odnosi się do sytuacji, w których znane są wartości zarówno zmiennych zależnych, jak i niezależnych. Na podstawie wartości tych zmiennych budowany jest model, który przewiduje wartości zmiennych zależnych dla nowych przypadków.

Do rozwiązywania problemów predykcyjnych można wykorzystać szereg metod data mining, takich jak: drzewa decyzyjne, sieci neuronowe, metodę wektorów wspierających (SVM), MARS czy metodę k-najbliższych sąsiadów. Dostępne są jednak także tradycyjne metody, takie jak regresja wieloraka dla problemów regresyjnych czy regresja logistyczna bądź analiza dyskryminacyjna, dla problemów klasyfikacyjnych. Zasadniczą różnicą między tymi dwiema grupami metod jest zdolność metod data mining do modelowania nieliniowych zależności oraz brak założeń formalnych co do rozkładów analizowanych zmiennych. Bardzo często również w tradycyjnym statystycznym podejściu nie interesuje nas, czy model skutecznie prognozuje (co jest kluczowe w data mining), tylko raczej czy oceny parametrów modelu są istotnie różne od zera i jaki procent zmienności prognozowanej zmiennej tłumaczy zbudowany model.

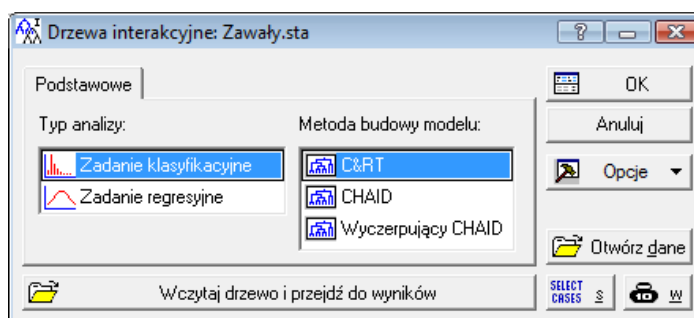


Wyszukiwanie reguł na przykładzie budowy modelu klasyfikacyjnego

Wykorzystanie technik data mining przedstawimy na przykładzie pliku *Zawały.sta* zawierającego informacje o wybranych parametrach biochemicznych oraz klinicznych zebranych dla pacjentów wraz z informacją, czy u danego pacjenta wystąpił zawał czy też nie. Do analizy wykorzystamy drzewa klasyfikacyjne w celu interakcyjnego wyszukiwania czynników i interakcji wpływających na ryzyko wystąpienia zawału.

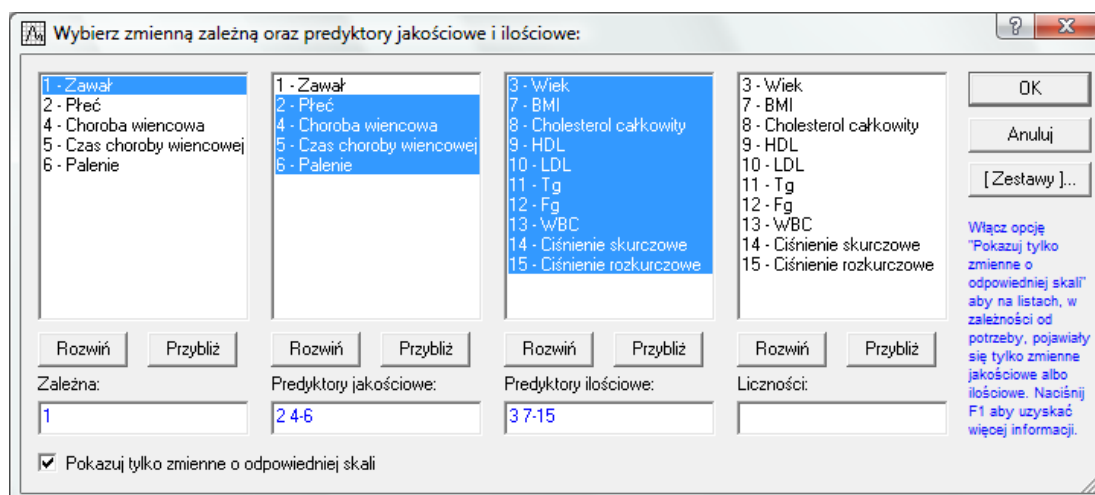
Analizowany zbiór zawiera 15 zmiennych. Zmienna *Zawał* informuje o fakcie wystąpienia zawału, natomiast kolejne 14 zmiennych to informacje o potencjalnych czynnikach wpływających na ryzyko jego wystąpienia.

W celu uruchomienia analizy z menu *Data Mining* wybieramy opcję *Drzewa Interakcyjne (C&RT, CHAID)*, a następnie określamy typ zadania jako *Zadanie klasyfikacyjne* (chcemy przewidywać zmienną *Zawał*, która jest zmienną jakościową), natomiast metodą, jakiej użyjemy do analizy, będą drzewa CART (*C&RT*).



Rys. 3. Drzewa interakcyjne – okno wstępnej specyfikacji analizy.

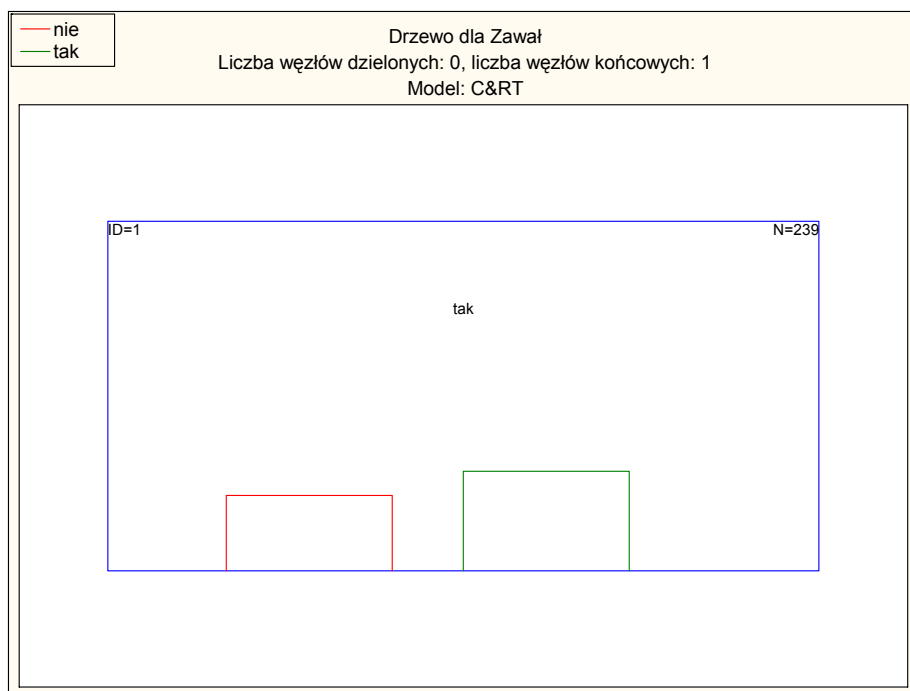
Po zatwierdzeniu wstępnych ustawień analizy w kolejnym kroku musimy określić zmienne, jakie będziemy analizować. Klikamy przycisk *Zmienne* znajdujący się na karcie *Podstawowe* i w oknie wyboru zmiennych wskazujemy zmienną *Zawał* jako zmienną zależną, zmienne *Płeć*, *Choroba wieńcowa*, *Czas choroby wieńcowej* oraz *Palenie* jako predyktory jakościowe, pozostałe zmienne określamy jako predyktory ilościowe.



Rys. 4. Wybór zmiennych do analizy.



Pozostałe parametry analizy pozostawiamy na domyślnych ustawieniach i przechodzimy do okna *Wyniki*. W tym momencie rozpoczynamy interakcyjne budowanie modelu oraz identyfikowanie czynników istotnie wpływających na ryzyko wystąpienia zawału.



Rys. 5. Węzeł macierzysty drzewa klasyfikacyjnego.

Na początku analizy nasze drzewo składa się jedynie z węzła macierzystego (rys. powyżej), który nie zawiera jeszcze żadnych węzłów potomnych. Widzimy, że w zbiorze danych występuje podobna proporcja osób z zawałem (słupek po prawej) co osób bez zawału (słupek po lewej). W trakcie analizy określimy podziały drzewa, które pozwolą wyodrębnić grupy w jak największym stopniu jednorodne ze względu na fakt wystąpienia zawału.

Predykcja dla węzła 1 (Zawały.sta) Model: C&RT		
	Typ podziału	Poprawa
Cholesterol całkowity	Automatycznie	0.123742
LDL	Automatycznie	0.113922
Czas choroby wieńcowej	Automatycznie	0.063157
Choroba wieńcowa	Automatycznie	0.062465
Ciśnienie skurczowe	Automatycznie	0.014151
HDL	Automatycznie	0.012693
BMI	Automatycznie	0.011023
Palenie	Automatycznie	0.010770
WBC	Automatycznie	0.010490
Fg	Automatycznie	0.008314
Wiek	Automatycznie	0.006863
Tg	Automatycznie	0.002278
Ciśnienie rozkurczowe	Automatycznie	0.001444
Płeć	Automatycznie	0.000130

Rys. 6. Ranking predyktorów.



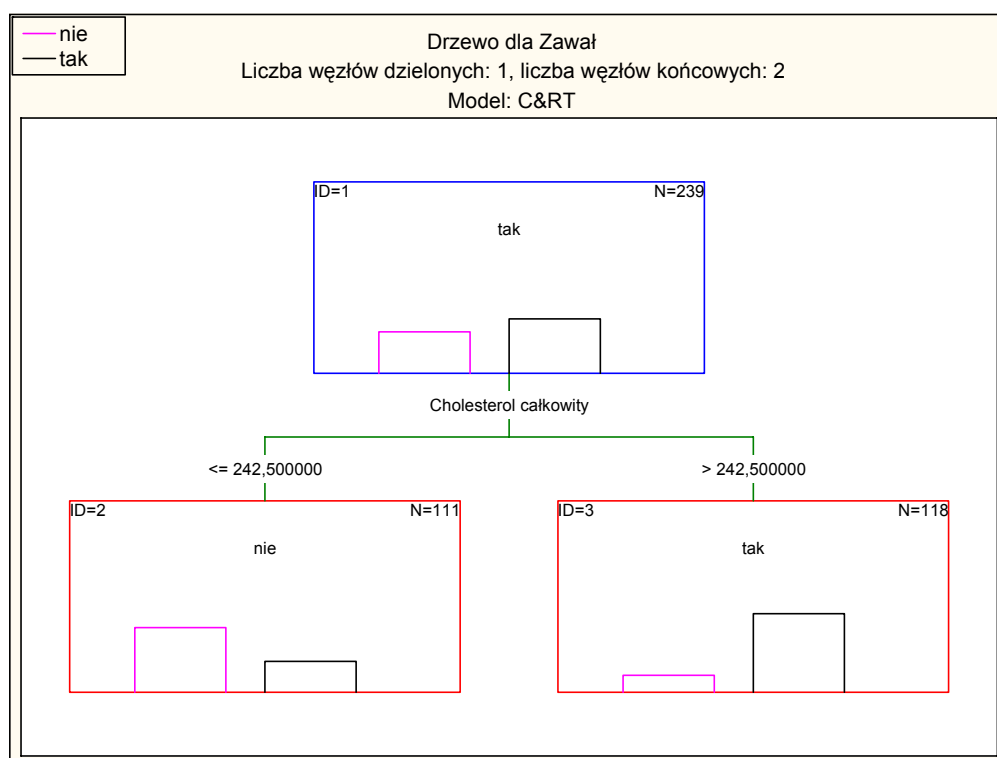
W pierwszej kolejności musimy wybrać zmienną, której użyjemy do pierwszego podziału. By móc to zrobić, warto wspomóc się rankingiem predyktorów dostępnym po naciśnięciu przycisku *Stat. Predyktorów*.

Widzimy, że najsilniejszym czynnikiem ryzyka jest *Cholesterol całkowity* i właśnie ta zmienna posłuży nam do pierwszego podziału.



Rys. 7. Określenie punktu podziału.

Klikamy przycisk *Określanie podziałów* i w wyświetlonym oknie widzimy, że algorytm sugeruje nam wprowadzenie podziału na poziomie 242,5. Akceptujemy ten poziom, klikając *Buduj*, co pozwoli wprowadzić nasz podział do modelu drzewa.

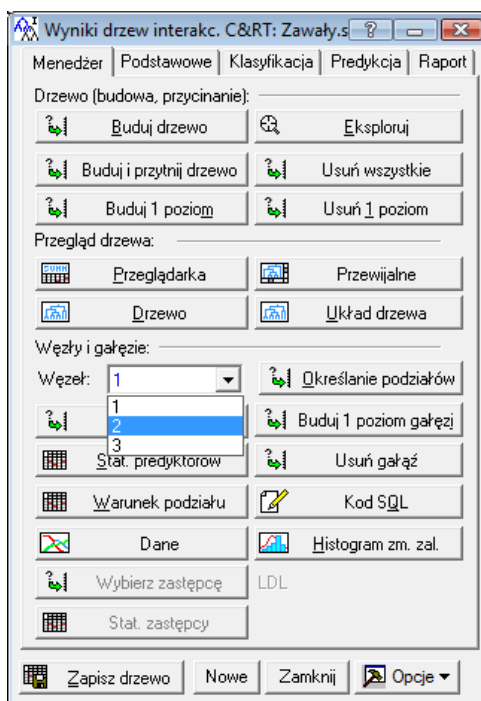


Rys. 8. Drzewo klasyfikacyjne z jednym podziałem.



Po wprowadzeniu podziału widzimy, że powstałe dwa węzły znacznie różnią się między sobą jeśli chodzi o rozkład zmiennej *Zawał*. Do lewego węzła trafiły osoby, których poziom cholesterolu całkowitego był mniejszy bądź równy wartości 242,5. W węźle tym możemy zauważyć znaczną przewagę osób bez zawału. W węźle prawym z kolei mamy osoby, których poziom cholesterolu był większy od 242,5, zdecydowaną przewagę mają w tym węźle osoby z zawałem.

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*.

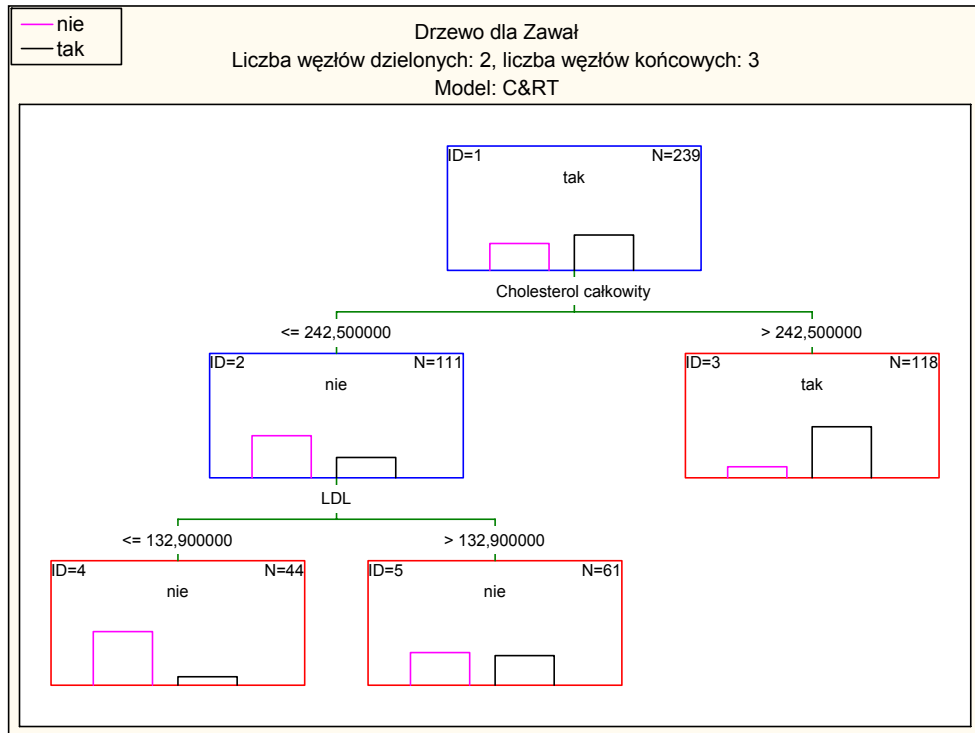


Rys. 9. Wyniki drzew interakcyjnych.

Tym razem najlepszą zmienną będącą podstawą podziału jest zmienna *LDL*. Program proponuje określić punkt podziału na poziomie 132,9. Akceptujemy ten punkt podziału, klikając *Buduj*.¹

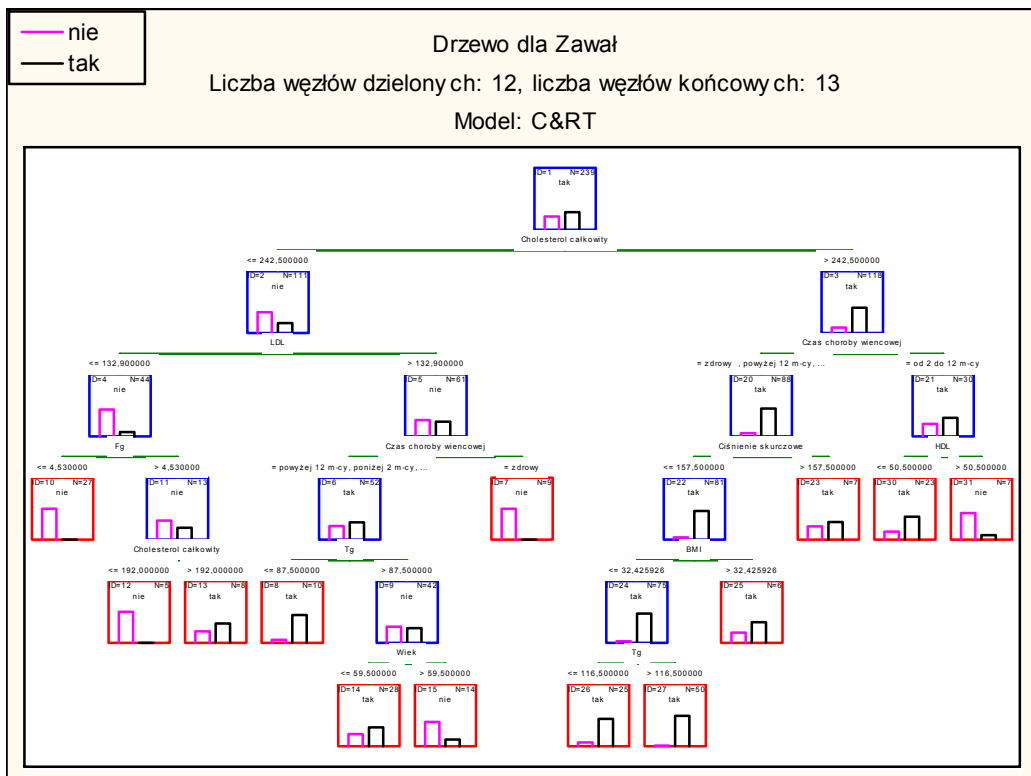
W wyniku podziału otrzymaliśmy nowe drzewo, które podzieliło zbiór danych na trzy grupy ryzyka. Jeśli dana osoba ma poziom LDL mniejszy bądź równy od 132,9 oraz jej poziom cholesterolu całkowitego jest mniejszy bądź równy 242,5, wtedy trafia do węzła 4 – czyli możemy ją zaliczyć do grona osób o najniższym ryzyku wystąpienia zawału serca. Osoby z węzła 5 to neutralne ryzyko, natomiast w węźle 3 znajdują się osoby o najwyższym ryzyku wystąpienia zawału.

¹ Korzystając z wiedzy eksperckiej, moglibyśmy zmienić zarówno zmienną służącą do podziału, jak też i poziom wprowadzanego podziału.



Rys. 10. Drzewo klasyfikacyjne z dwoma podziałami.

Oczywiście każdy z utworzonych węzłów końcowych możemy dalej dzielić w celu określenia bardziej jednorodnych węzłów. Poza zaprezentowanym powyżej ręcznym budowaniem drzewa zawsze mamy możliwość zbudowania go w sposób automatyczny.



Rys. 11. Drzewo decyzyjne zbudowane w sposób automatyczny.



Po naciśnięciu przycisku *Buduj drzewo* otrzymujemy zbudowany automatycznie dużo bardziej skomplikowany model składający się z 13 węzłów końcowych (zob. rys. 11).

Dla wszystkich utworzonych węzłów końcowych możemy podać łatwą w interpretacji regułę opisującą osoby przypisane do poszczególnych grup ryzyka (podobnie jak uczyniliśmy to z węzłem 4).

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 liczności, 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.

Przykład wykorzystania przestrzeni roboczej STATISTICA Data Miner do rozwiązania problemu regresyjnego

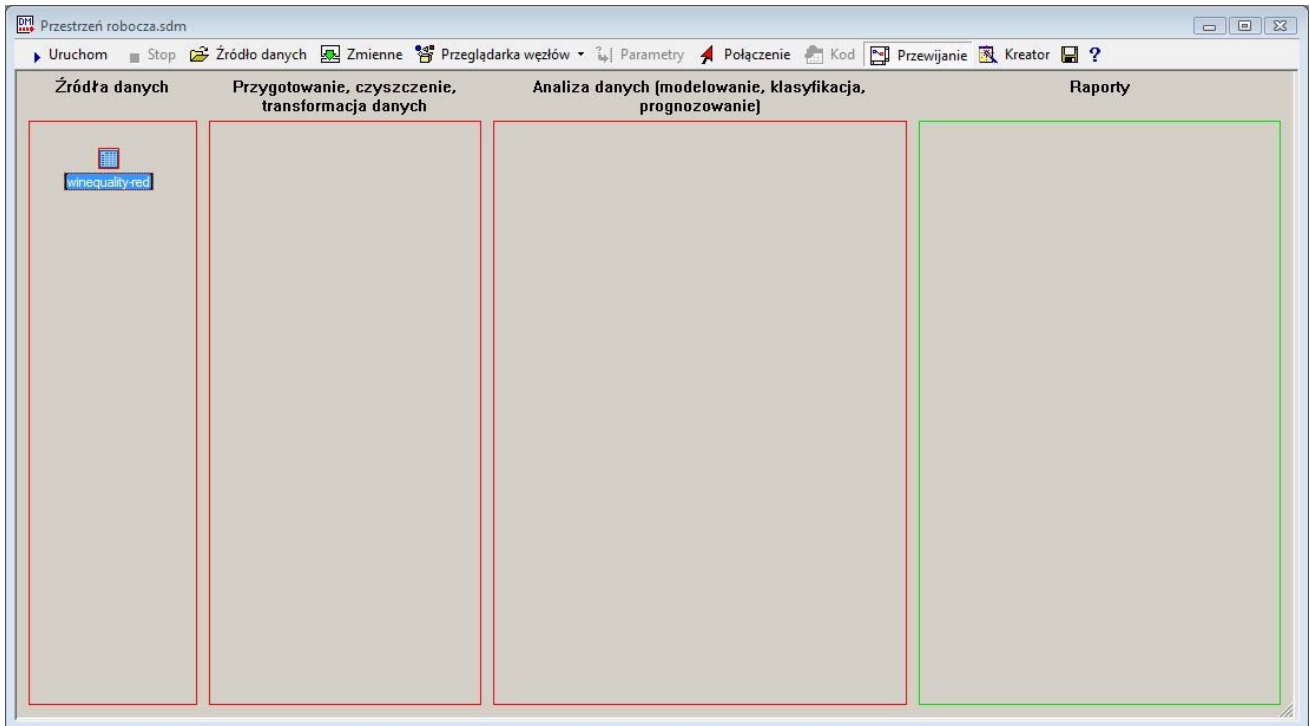
W kolejnym przykładzie naszym celem będzie zbudowanie i porównanie modeli przewidyujących jakość czerwonego wina pochodzącego z jednego z regionów Portugalii na podstawie jego parametrów, takich jak: zawartość alkoholu, kwasowość czy zawartość siarczynów.

Dysponujemy zbiorem danych zawierającym 1599 przypadków, każdy przypadek opisuje jeden rodzaj wina. Wina opisane są za pomocą 11 zmiennych, dwunasta zmienna informuje o jakości danego rodzaju wina. Naszym zadaniem będzie zbudowanie modelu zależności pomiędzy parametrami a jakością wina. Ponieważ jakość wina wyrażona jest na skali ilościowej, nasz problem zaliczymy do grupy zadań regresyjnych.

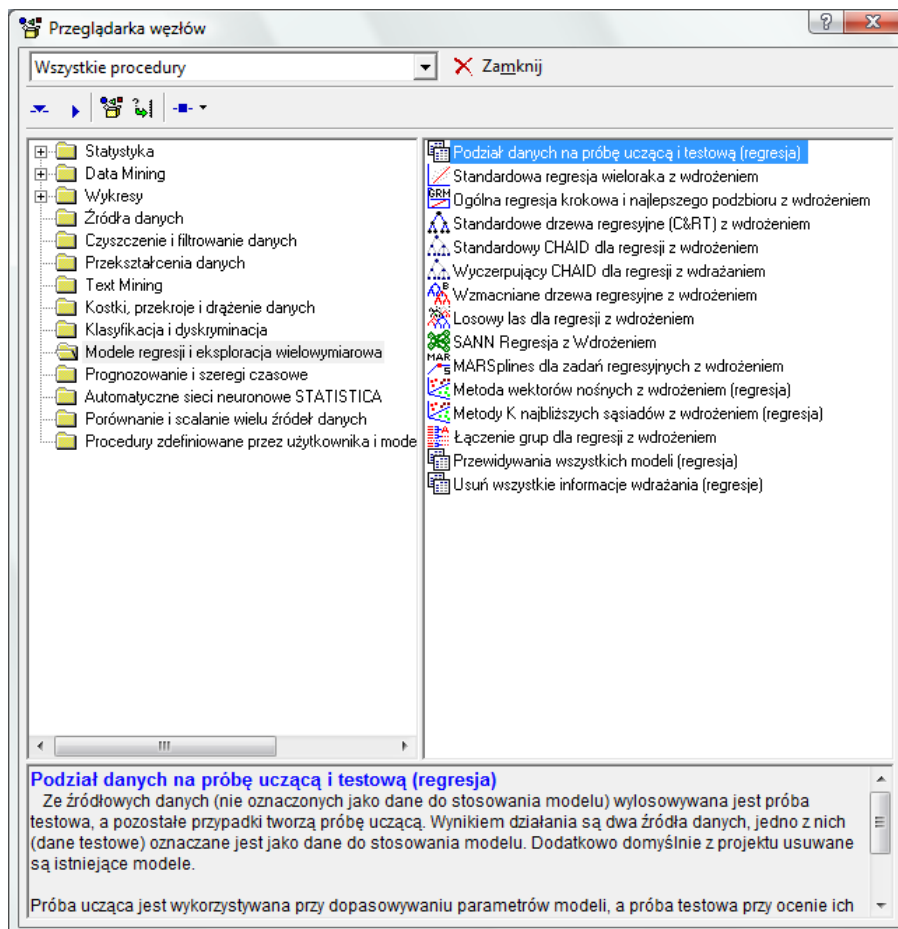
Do zbudowania modeli predykcyjnych wykorzystamy tym razem przestrzeń roboczą programu *STATISTICA Data Miner*, która umożliwia przygotowanie kompletnego projektu analitycznego od zrozumienia i przygotowania danych, poprzez modelowanie, aż po ocenę zbudowanych modeli.

Aby uruchomić przestrzeń roboczą, z menu *Data mining* wybieramy opcję *Przestrzeń robocze*, a następnie polecenie *Wszystkie procedury*. W wyświetlonej przestrzeni roboczej klikamy przycisk *Źródło danych* i wybieramy plik *winequality-red.sta*, a następnie wybieramy zmienne do analizy. Zmienną *Jakość* wskazujemy jako zmienną zależną ilościową, natomiast wszystkie pozostałe zmienne definiujemy jako predyktory ilościowe.

W kolejnym kroku klikamy przycisk *Przeglądarka węzłów*, dzięki czemu w wyświetlonym oknie przeglądarki węzłów możemy wybrać i wstawić do przestrzeni roboczej potrzebne moduły analityczne. By nie wprowadzać do naszego przykładu niepotrzebnych komplikacji, ominiemy najbardziej żmudny etap analizy, jakim jest czyszczenie i przygotowywanie danych, i przyjmiemy (zresztą zgodnie z prawdą), że nasze dane są odpowiedniej jakości.



Rys. 12. Przestrzeń robocza programu *STATISTICA Data Miner*.



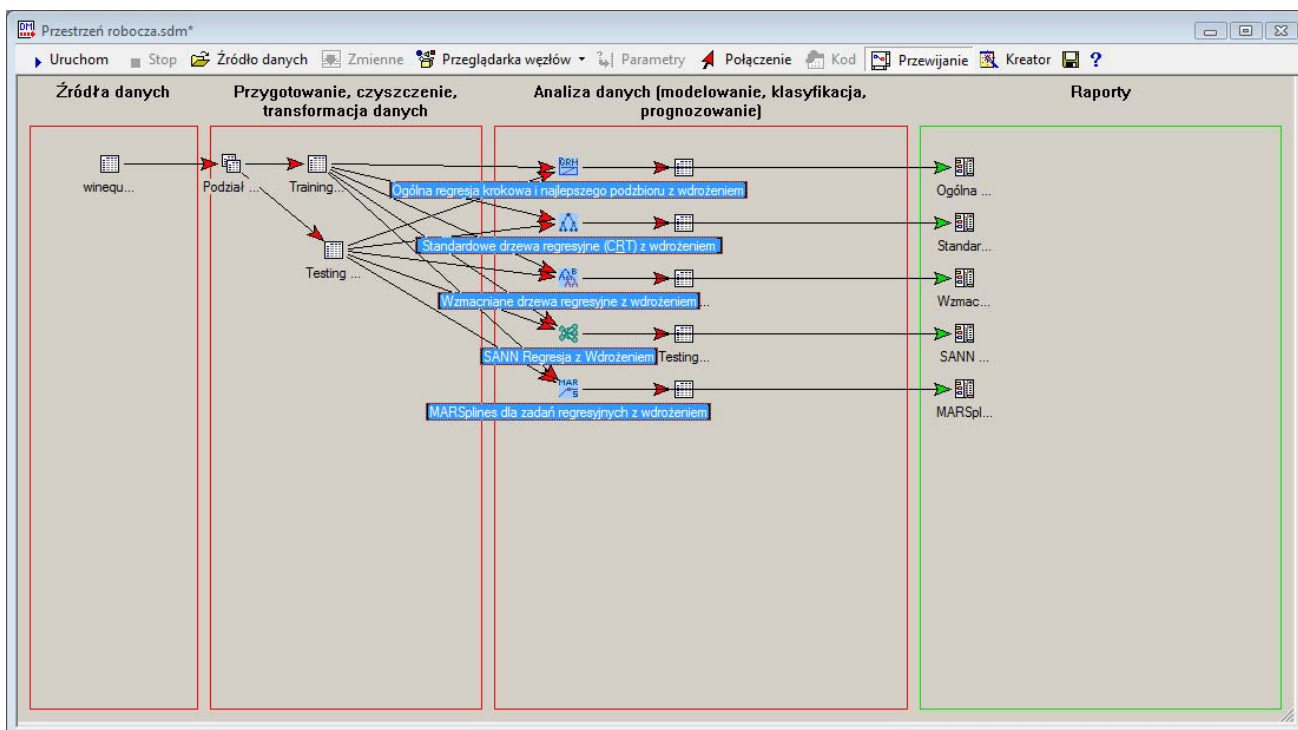
Rys. 13. Przeglądarka węzłów.



Zanim przejdziemy do samego modelowania, podzielimy nasz zbiór na dwa podzbiory. Pierwszy z nich pełni rolę próby uczącej – na jego podstawie szacować będziemy parametry modeli. Drugi pełni rolę próby testowej i będzie służył do oceny zdolności modeli do generalizacji. Aby podzielić nasz zbiór na dwa podzbiory, w przeglądarce węzłów przechodzimy do katalogu *Modele regresji i eksploracja wielowymiarowa* i klikamy dwukrotnie na węzeł *Podział danych na próbę uczącą i testową (regresja)*, co spowoduje, że wybrany węzeł zostanie wstawiony do przestrzeni roboczej.

Następnie zmieniamy domyślne parametry węzła, określając, że do zbioru testowego powinno trafić około 30% przypadków. Po określeniu parametrów losowania uruchamiamy projekt klikając przycisk *Uruchom*.

Do budowy modelu użyjemy równolegle szeregu metod analitycznych. W przeglądarce węzłów kolejno klikamy na węzły: *Ogólna regresja krokowa i najlepszego podzbioru z wdrożeniem*, *Standardowe drzewa regresyjne (C&RT) z wdrożeniem*, *Wzmacniane drzewa regresyjne z wdrożeniem*, *SANN Regresja z wdrożeniem* (sieci neuronowe), *MARSplines dla zadań regresyjnych z wdrożeniem*. Po umieszczeniu wybranych węzłów w przestrzeni roboczej łączymy je ze zbiorem uczącym oraz testowym, a następnie zmieniamy parametry niektórych metod. Dla regresji krokowej określamy metodę budowy modelu jako Krokowa wsteczna, w przypadku drzew regresyjnych włączamy opcję V-krotny sprawdzian krzyżowy, aby uniknąć nadmiernej rozbudowy struktury modelu. W przypadku metody MARSplines zwiększamy rząd interakcji do poziomu 2. Po określeniu dodatkowych parametrów klikamy przycisk *Uruchom*, aby rozpocząć proces budowy modeli.



Rys. 14. Przestrzeń robocza ze wstawionymi modułami analitycznymi oraz raportami z budowy modeli.

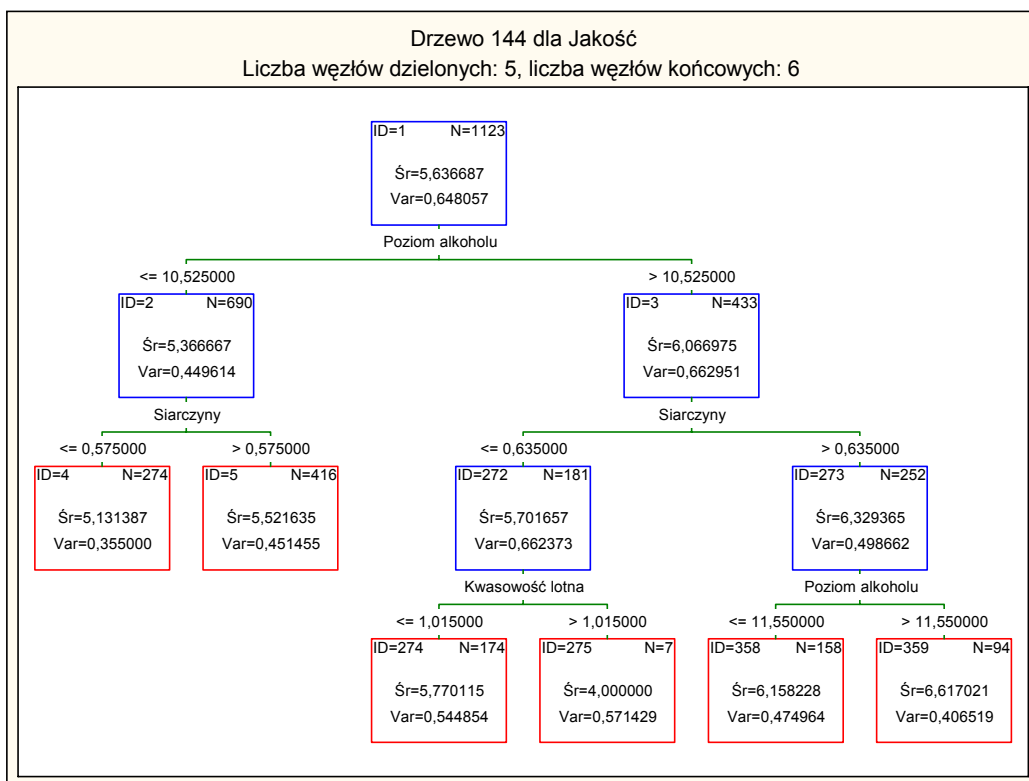


Po zbudowaniu zestawu modeli w obszarze *Raporty* dostępne są podsumowania dla poszczególnych metod. Przykładowo klikając na raport dla regresji wielorakiej, możemy ocenić wartości ocen parametrów regresji czy też zbadać przebieg doboru optymalnego zestawu parametrów. Widzimy, że proces doboru parametrów trwał 6 iteracji – z modelu wyeliminowanych zostało 5 zmiennych, które zostały uznane za nieistotne.

	F do usunięc.	P do usunięc.	F do wprowadz.	P do wprowadz.	Efekt (stan)
Siarczyny	40.796	0.000			W modelu
Kwasowość lotna	74.372	0.000			W modelu
pH	6.999	0.008			W modelu
Poziom alkoholu	178.965	0.000			W modelu
Chlorki	14.519	0.000			W modelu
Całkowity dwutlenek siarki	18.096	0.000			W modelu
Wolny dwutlenek siarki			2.978	0.085	Poza
Cukier osadowy			1.063	0.303	Poza
Kwas cytrynowy			1.031	0.310	Poza
Kwasowość stała			0.001	0.975	Poza
Gęstość			0.007	0.932	Poza

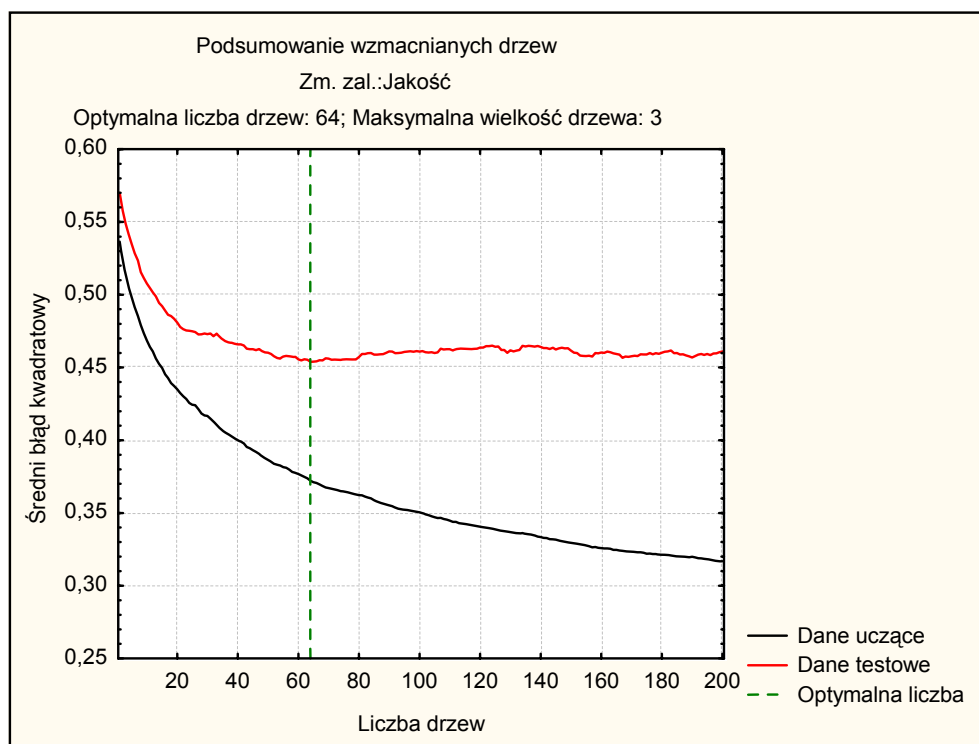
Rys. 15. Fragment raportu budowy modelu regresji wielorakiej.

W przypadku drzew regresyjnych możemy między innymi zbadać strukturę drzewa, by ocenić, jakie zmienne weszły do modelu. Widzimy, że w modelu uwzględniono jedynie trzy zmienne: *Poziom alkoholu*, *Siarczyny* oraz *Kwasowość lotna*.



Rys. 16. Model drzew regresyjnych C&RT.

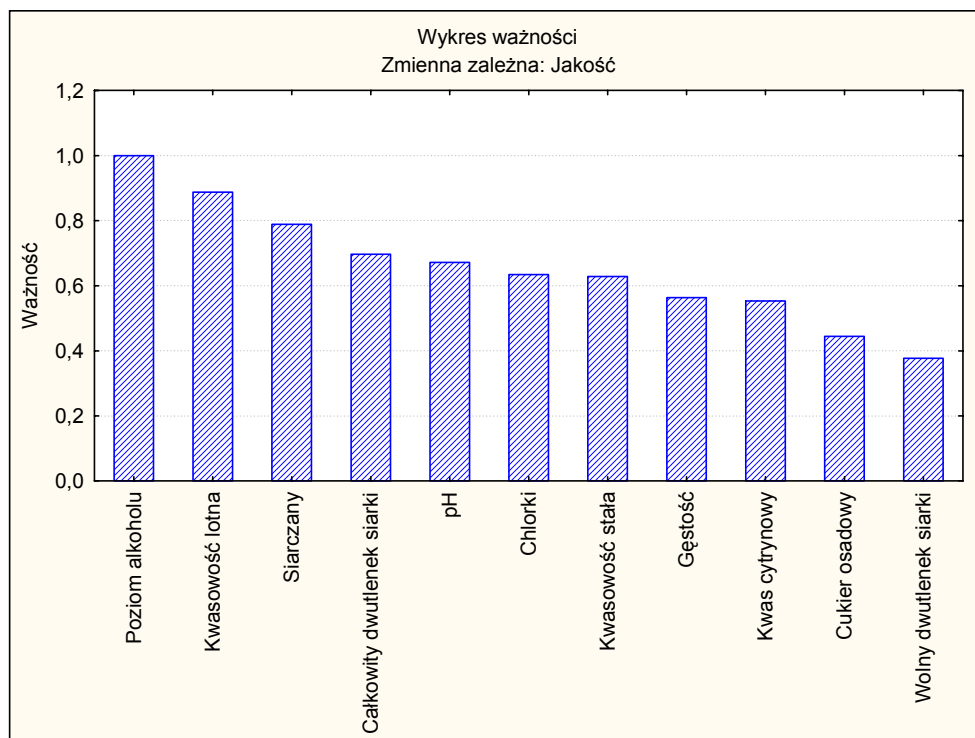
W przypadku drzew wzmocnianych model składa się z wielu prostych modeli budowanych na podpróbach wylosowanych ze zbioru uczącego po wcześniejszym określeniu wag przypadków, które zwiększają prawdopodobieństwo wylosowania do kolejnego zbioru tych przypadków, które generowały największy błąd. Zwykle model składa się z kilkuset prostych modeli, które działając wspólnie dają uśrednioną prognozę. Analizując przebieg budowy modelu drzew wzmocnianych, widzimy, że zwiększanie liczby drzew zawsze prowadziło do spadku błędu na danych uczących (błąd aproksymacji), natomiast błąd dla danych testowych (błąd generalizacji) malał jedynie w początkowej fazie budowy modelu. Począwszy od 65 drzew zaczął rosnąć, co zwykle sugeruje nadmierne dopasowywanie się modelu do danych. Po zbudowaniu 200 drzew program przywrócił stan modelu do 64 drzew, czyli punktu, w którym błąd generalizacji był najmniejszy.



Rys. 17. Przebieg budowy wzmocnianych drzew.

Dodatkowo dla drzew wzmocnianych możemy wygenerować wykres ważności predyktorów (rys. 18 poniżej). Widzimy, że trzy najważniejsze predyktory pokrywają się ze zmiennymi, które weszły do modelu drzew C&RT.

Na koniec naszej analizy porównamy jakość zbudowanych modeli. W tym celu z przeglądarki węzłów wybieramy moduł *Przewidywania wszystkich modeli (regresja)* i łączymy go ze zbiorem testowym. Po jego uruchomieniu otrzymujemy węzeł, w którym zawarte są przewidywania wszystkich modeli dla win ze zbioru testowego oraz dodatkowo przewidywanie uśrednione – przygotowanie na podstawie wszystkich modeli.



Rys. 18. Ranking ważności zmiennych dla wzmacnianych drzew.

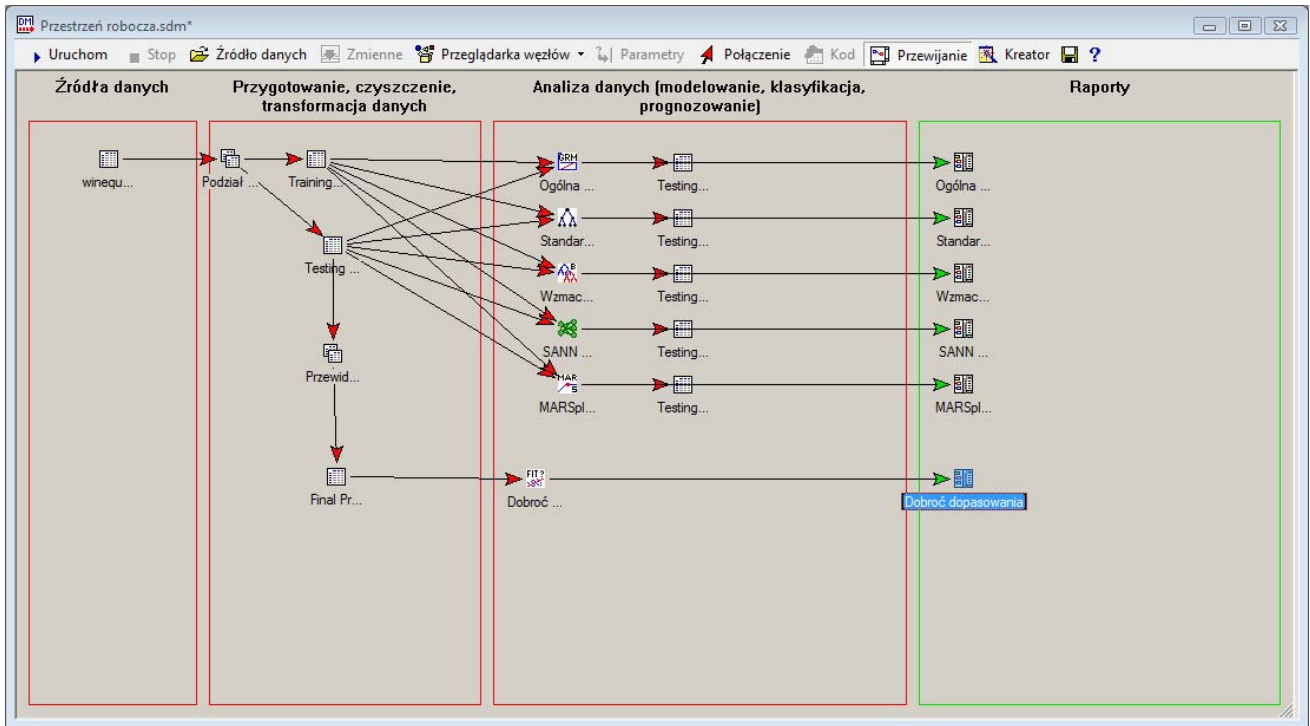
Do tak przygotowanego zbioru podłączamy węzeł służący do oceny dobroci dopasowania – z przeglądarki węzłów wybieramy węzeł *Dobroć dopasowania* (katalog *Data mining->Dobroć dopasowania*) i uruchamiamy go, uzyskując poniższy raport.

	Regresja wieloraka	Drzewa C&RT	MARS	Drzewa wzmacniane	Sieci neuronowe	Uśrednione przewidywanie
Średnia kwadratów reszt	0.386	0.432	0.396	0.405	0.360	0.361
Średni błąd bezwzględny	0.486	0.518	0.487	0.505	0.465	0.467

Rys. 19. Porównanie dopasowania modeli regresji.

Widzimy, że najlepszym modelem z punktu widzenia średniego błędu bezwzględnego oraz średniej kwadratów reszt jest model sieci neuronowej. Dobre przewidywania daje również uśredniony model oraz model zbudowany na podstawie regresji wielorakiej.

Zbudowany projekt data mining możemy następnie zapisać w postaci pliku i posłużyć się nim do przewidywania nowych przypadków.



Rys. 20. Końcowy kształt projektu data mining.

Literatura

1. Berry M., Linoff G., *Data Mining Techniques for Marketing, Sales and Customer Support*, John Wiley & Sons, Inc, New York 1997.
2. Berry M., Linoff G., *Mastering Data Mining. The Art and Science of Customer Relationship Management*, John Wiley & Sons, Inc, New York 2000.
3. Goodman A., Kamath C., Kumar V. *Data analysis in the 21st Century, Statistical Analysis and Data Mining*, John Wiley & Sons, Inc. New York, NY, USA, 2008.
4. Guidici P. *Applied Data Mining Statistical Methods for Business and Industry*, John Wiley & Sons, Inc, 2003.
5. Tadeusiewicz R. *Data mining jako szansa na relatywnie tanie dokonywanie odkryć naukowych poprzez przekopywanie pozornie całkowicie wyeksploatowanych danych empirycznych*, [w:] Materiały na seminarium „Zastosowania statystyki i data mining w badaniach naukowych”, StatSoft Polska, Kraków 2006.