



## PORÓWNANIE MODELI KLASYFIKACYJNYCH NA PRZYKŁADZIE DANYCH DOTYCZĄCYCH MŁODYCH PACJENTÓW Z LAMINEKTOMIĄ KRĘGOSŁUPA

*dr Janusz Wątroba<sup>2</sup>*

Prezentowany przykład ilustruje zastosowanie dwóch różnych metod budowania modeli w przypadku zagadnienia klasyfikacyjnego. Zaprezentowane zostaną dwie techniki klasyfikacyjne: *ogólna analiza dyskryminacyjna* i *ogólne modele drzew klasyfikacyjnych i regresyjnych*.

### Wprowadzenie

Przykładowe dane dotyczą młodych pacjentów z laminektomią kręgosłupa w odcinku piersiowym i lędźwiowym. Chorzy ci byli poddani operacji kręgosłupa z powodu występowania nowotworu, wad wrodzonych lub wad rozwojowych. Celem badań było określenie faktycznego występowania i natury deformacji kręgosłupa po zabiegu operacyjnym, ocena istotności wpływu takich czynników jak wiek w momencie operacji oraz liczba i położenie zdeformowanych kręgów. Dane do przykładu zostały zaczerpnięte z książki, której autorami są Hastie i Tibshirani [2].

Warto podkreślić w tym miejscu, że z podobnie postawionym zagadnieniem możemy się spotkać także w wielu innych dziedzinach. Budowa modeli może przy tym dotyczyć zarówno zagadnień o charakterze poznawczym (kiedy chodzi nam o poznanie struktury czynników wpływających na określone zjawisko lub proces), jak i w zagadnieniach praktycznych (kiedy chodzi nam np. o wspomaganie procesów decyzyjnych). Z sytuacją taką moglibyśmy mieć do czynienia np.:

- ♦ w badaniach medycznych (przy ocenie czynników wpływających na zapadalność na daną chorobę lub przy ocenie skuteczności określonej metody leczenia),
- ♦ w badaniach pedagogicznych (przy ocenie czynników determinujących osiągnięcia szkolne),

---

<sup>2</sup> StatSoft Polska Sp. z o.o.



- ◆ w zagadnieniach bankowych (przy ocenie czynników wpływających na spłacalność kredytu),
- ◆ w zagadnieniach ubezpieczeniowych (przy ocenie czynników determinujących ryzyko powstania szkody w ubezpieczeniach komunikacyjnych).

Użyty w niniejszym przykładzie plik danych zawiera następujące zmienne:

- ◆ **Kifoza**; zmienna jakościowa zawierająca informację o występowaniu kifozy u każdego z badanych pacjentów (1 - występuje, 2 - nie występuje),
- ◆ **Wiek [m-ce]**; zmienna ilościowa określająca wiek pacjenta (wyrażony w miesiącach) w momencie przeprowadzenia zabiegu operacyjnego,
- ◆ **Początkowy kręę**; zmienna ilościowa oznaczająca numer początkowego kręęu odcinka kręęoslupa, który został poddany zabiegowi operacyjnemu,
- ◆ **Liczba kręęów**; zmienna ilościowa określająca liczbę kręęów, które zostały objęte zabiegiem operacyjnym.

Poniżej zamieszczono fragment arkusza danych programu *STATISTICA*, który zawiera wykorzystywane dane.

Dane dotyczą młodych pacjentów z laminectomią kręęoslupa w odcinku piersiowym i lędźwiowym. Chorzy byli poddani operacji kręęoslupa ze względu na występowanie nowotworu, wad wrodzonych lub wad rozwojowych. Celem badań było określenie rzeczywistego występowania i natury deformacji kręęoslupa po zabiegu operacyjnym oraz ocena istotności wpływu takich czynników jak wiek w momencie operacji oraz liczba i położenie zdeformowanych kręęów.				
	1 Kifoza	2 Wiek [m-ce]	3 Początkowy kręę	4 Liczba kręęów
1	nie występuje	71	5	3
2	nie występuje	158	14	3
3	występuje	128	5	4
4	nie występuje	2	1	5
5	nie występuje	1	15	4
6	nie występuje	1	16	2
7	nie występuje	61	17	2
8	nie występuje	37	16	3
9	nie występuje	113	16	2
10	występuje	59	12	6
11	występuje	82	14	5

Głównym celem prezentowanego przykładu będzie budowa modeli, pozwalających na prawidłową klasyfikację przypadków występowania lub niewystępowania kifozy w oparciu o wartości zawarte w pliku danych zmiennych objaśniających (predyktorów), o charakterze ilościowym.

Drugim celem niniejszego przykładu jest praktyczna prezentacja sposobu przeprowadzania bardziej złożonych analiz z użyciem narzędzi statystycznych zawartych w programie *STATISTICA Data Miner*.



## Wstępna charakterystyka analizowanych zmiennych

Na wstępie analizy zobaczymy jak przedstawia się rozkład wartości zmiennej **Kifoza**. W zamieszczonej poniżej tabelce możemy zobaczyć, że zdecydowaną większość (ponad 78%) stanowili chorzy, u których po zabiegu operacyjnym nie stwierdzono występowania kifozy.

Kategoria	Tabela licznosci: Kifoza	
	Licznosc	Procent
wystepuje:	18	21,68675
nie wystepuje:	65	78,31325
Braki	0	0,00000

W kolejnym kroku analizy zobaczymy, czy grupy pacjentów wyróżnione za pomocą kategorii zmiennej **Kifoza** różnią się pod względem średnich wartości zmiennych ilościowych. W tym celu przeprowadzimy tzw. *analizę przekrojową*. Wyniki analizy, w postaci podstawowych statystyk opisowych dla kolejnych zmiennych, zawierają poniższe table:

Kifoza	Wiek [m-ce]	Wiek [m-ce]	Wiek [m-ce]	Wiek [m-ce]	Wiek [m-ce]
	Średnie	N	Odch.std	Minimum	Maksimum
wystepuje	93,05556	18	43,13927	12,00000	157,0000
nie wystepuje	83,93846	65	64,23221	1,00000	243,0000
Ogół grp	85,91566	83	60,16830	1,00000	243,0000

Wyniki zamieszczone w tabeli pokazują, że pacjenci, u których stwierdzono pooperacyjne deformacje kręgosłupa, byli przeciętnie o około 9 miesięcy starsi w momencie przeprowadzenia zabiegu. Jednocześnie rozrzut wieku, mierzony odchyleniem standardowym, był w tej grupie zdecydowanie mniejszy ( $s = 43,14$ ).

Następna tabela zawiera analogiczne wyniki dla zmiennej **Początkowy kręę**.

Kifoza	Początkowy kręę	Początkowy kręę	Początkowy kręę	Początkowy kręę	Początkowy kręę
	Średnie	N	Odch.std	Minimum	Maksimum
wystepuje	7,00000	18	4,338609	1,000000	14,00000
nie wystepuje	12,53846	65	4,430283	1,000000	18,00000
Ogół grp	11,33735	83	4,949198	1,000000	18,00000

Widać dość wyraźnie (biorąc pod uwagę wartości średnie), że zakres kręęów objętych zabiegiem u pacjentów z kifozą obejmował kręęgi położone wyżej (przeciętnie o około 5-6 kręęów). Tym razem rozrzut wartości jest w obu grupach zbliżony (odchylenie standardowe wynosi około 4,4).

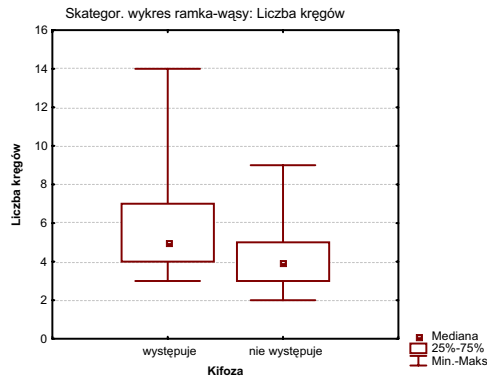
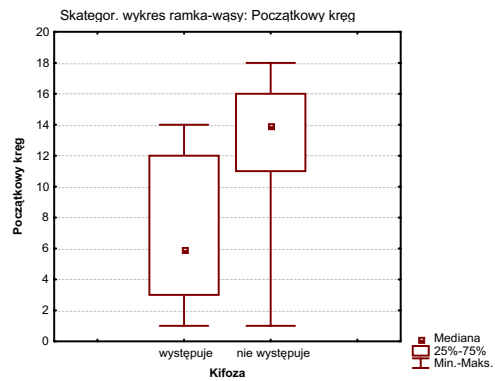
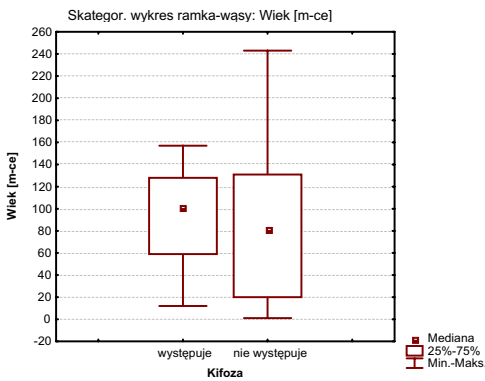
Wyniki zamieszczone w kolejnej tabeli (patrz poniżej) pokazują ponadto, że u pacjentów, u których stwierdzono występowanie kifozy, zabieg operacyjny obejmował przeciętnie



o około 2 kręgi szerszy zakres (przy blisko dwukrotnie większym rozrzucie wartości) niż u pacjentów, u których kifoza nie występowała.

Tabela dwudzielcza statystyk opisowych (Kifoza)					
N=83 (Zmienne zależne nie zawierają BD)					
Kifoza	Liczba kręgów Średnie	Liczba kręgów N	Liczba kręgów Odch. std	Liczba kręgów Minimum	Liczba kręgów Maksimum
występuje	5,666667	18	2,765332	3,000000	14,00000
nie występuje	3,815385	65	1,498878	2,000000	9,00000
Ogół grp	4,216867	83	1,981919	2,000000	14,00000

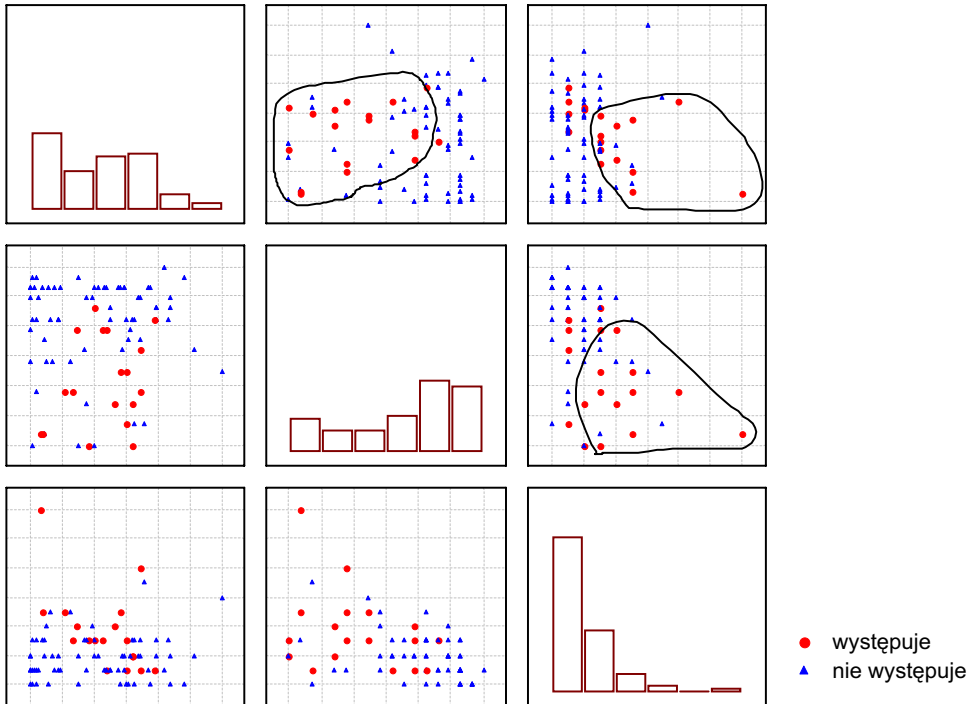
Dla porównania rozkładu wartości analizowanych zmiennych ilościowych poniżej zamieszczono odpowiednie wykresy typu ramka-wąsy (na wykresach zamieszczono wartości median, kwartyli oraz wartość minimalną i maksymalną).



Przed przystąpieniem do budowania odpowiednich modeli, umożliwiających poprawną klasyfikację pacjentów, przyjrzymy się jeszcze powiązaniom występującym pomiędzy analizowanymi zmiennymi objaśniającymi. W tym celu w programie *STATISTICA* utworzono macierzowy wykres rozrzutu, na którym przynależność do kategorii zmiennej **Kifoza** została oznaczona różnymi znacznikami punktów.

Analiza przedstawionego poniżej wykresu pozwala sądzić, że powiązania pomiędzy zmiennymi umożliwią utworzenie dobrych modeli klasyfikacyjnych. Na wykresie widać bowiem, że punkty oznaczające poszczególne kategorie wykazują tendencję do grupowania się w oddzielnych obszarach wykresu.

Wykres macierzowy (Kifoza 4v\*83c)



Zasadnicza część analizy, w której będziemy się starali budować odpowiednie modele klasyfikacyjne, będzie prowadzona za pomocą programu *STATISTICA Data Miner*. W związku z tym poniżej został zamieszczony krótki opis budowy programu oraz charakterystyka sposobu pracy z programem.

## Struktura i interfejs użytkownika w programie *STATISTICA Data Miner*

*STATISTICA Data Miner* to kompletny zestaw narzędzi data mining, zaprojektowany tak, aby umożliwić łatwe i szybkie przeprowadzanie analizy danych i zastosowanie uzyskanych wyników we wspomaganie podejmowania decyzji. Narzędzia zaimplementowane w programie cechuje wysoka wydajność na wszystkich etapach wydobywania z danych użytecznej wiedzy; począwszy od pobierania danych z baz i hurtowni danych, a skończywszy na tworzeniu raportów. System został zoptymalizowany pod kątem wydajności analizy dużych zbiorów danych. Zawarte w nim wyrafinowane techniki analizy danych umożliwiają efektywne rozwiązywanie nawet najtrudniejszych problemów analitycznych.



*STATISTICA Data Miner* łączy w sobie zalety intuicyjnego, interaktywnego interfejsu z pełną programowalnością i szerokimi możliwościami dostosowywania do konkretnych zadań. Niektórzy użytkownicy systemu nie będą musieli w ogóle wychodzić poza interaktywny tryb pracy; inni jednak będą mogli wykorzystać ogromne możliwości obiektowo zorientowanego interfejsu programistycznego, wykorzystując skrypty Visual Basic.

Podstawowym elementem systemu *STATISTICA Data Miner* jest zbiór ponad 260 wysoce zoptymalizowanych, efektywnych i niezwykle szybkich procedur programu *STATISTICA*, wywoływanych za pomocą skryptów Visual Basic (dostępnych jako kody źródłowe), które są wykorzystywane do określania relacji pomiędzy procedurami (obiektami) i do sterowania ogólną logiką projektu (oraz „przepływem” danych). Ta elastyczna, dostosowywalna architektura udostępnia pełną funkcjonalność wszystkich procedur statystycznych i analitycznych dla środowiska data mining w postaci obiektów zawierających analizy. Skrypty (obiekty analizy) służą jako „pojemniki” lub szablony definiujące sposób przepływu danych w projekcie, podczas gdy rzeczywiste analizy numeryczne są przeprowadzane za pomocą niezwykle szybkich procedur analitycznych programu *STATISTICA*. Obiekty, które mogą być wykorzystywane w charakterze węzłów dla operacji „czyszczenia” i filtrowania danych oraz do analizy danych, są zawarte w *Przeglądarce węzłów*.

Tworzenie projektu data mining odbywa się w specjalnie zaprojektowanym obszarze okna programu. Obszar projektów data mining ma określoną strukturę i stanowi bardzo efektywne, przyjazne dla użytkownika środowisko analizy danych, w którym możemy przenosić i wzajemnie łączyć dane, analizy i wyniki przez proste przeciąganie ikon i strzałek symbolizujących połączenia. Możemy jednocześnie otwierać, modyfikować i uruchamiać dowolną liczbę obszarów projektów data mining, a także przeciągać węzły (obiekty) pomiędzy różnymi obszarami i przeglądarkami węzłów. Obszar projektów data mining został wstępnie podzielony na cztery panele:

- ♦ **Źródło danych.** W tym panelu określamy źródła danych, np. pliki danych programu *STATISTICA*, elementy obrazujące bazy danych przeznaczone do zdalnego przetwarzania na serwerach zewnętrznych czy też programy generujące dane automatycznie do zastosowania w procesie zaawansowanego modelowania.
- ♦ **Przygotowywanie, czyszczenie i przekształcanie danych.** Węzły występujące w tej części akceptują na wejściu jedno lub większą liczbę źródeł danych i tworzą jedno lub większą liczbę (odfiltrowanych, oczyszczonych i przekształconych) źródeł danych dla dalszych, bardziej „dogłębnych” analiz.
- ♦ **Analiza danych, modelowanie, klasyfikacja i prognozowanie.** Węzły występujące w tej części służą do przeprowadzania analizy danych.
- ♦ **Raporty.** W tej części obszaru projektów data mining umieszczane są wyniki poszczególnych analiz.

Tworzenie projektu data mining jest łatwe: w pierwszym kroku wybieramy źródło danych, w drugim kroku stosujemy wymagane operacje przygotowywania, oczyszczania



i przekształcania danych, w trzecim kroku łączymy wymagane przez nas analizy z oczyszczonymi danymi i w czwartym kroku przeglądamy lub publikujemy wyniki.

W kolejnej części zostaną zaprezentowane przykłady budowania modeli klasyfikacyjnych, kolejno przy użyciu *ogólnej analizy dyskryminacyjnej* oraz *ogólnych modeli drzew klasyfikacyjnych i regresyjnych*.

## **Budowa modelu klasyfikacyjnego przy użyciu ogólnej analizy dyskryminacyjnej**

Przy budowie pierwszego modelu wykorzystamy możliwości modułu *Ogólne modele analizy dyskryminacyjnej (GDA, General Discriminant Analysis)*. Zawarta w nim metoda analizy stanowi rozszerzenie możliwości klasycznej analizy dyskryminacyjnej, polegające na tym, że w charakterze zmiennych objaśniających mogą występować zarówno zmienne jakościowe, jak i ilościowe. Słowo „ogólna” występujące w nazwie metody podkreśla fakt, iż do zagadnienia analizy funkcji dyskryminacyjnej stosowany jest ogólny model liniowy (GLM, General Linear Model). W metodzie ogólnej analizy dyskryminacyjnej zagadnienie analizy funkcji dyskryminacyjnej zostało „przetworzone” do postaci ogólnego wielowymiarowego modelu liniowego, w którym analizowane zmienne zależne są zakodowanymi (zero-jedynkowo) wektorami, odzwierciedlającymi przynależność każdego z przypadków do określonej grupy.

Jedną z korzyści wynikających ze stosowania ogólnego modelu liniowego do zagadnienia analizy dyskryminacyjnej jest możliwość definiowania złożonych modeli dla zbioru predyktorów. I tak przykładowo, dla zbioru predyktorów ilościowych możemy zdefiniować model regresji wielomianowej, model powierzchni odpowiedzi, model regresji czynnikowej lub model regresji powierzchni odpowiedzi dla mieszaniny (bez wyrazu wolnego). Moduł GDA w programie *STATISTICA* rzeczywiście nie nakłada żadnych szczególnych ograniczeń na typ stosowanych predyktorów (jakościowych czy ilościowych) czy też typ definiowanego modelu. Jednakże w przypadku stosowania predyktorów jakościowych trzeba zachować pewną ostrożność.

Oprócz tradycyjnej, krokowej analizy dyskryminacyjnej dla predyktorów ilościowych, która jest dostępna w module *Analiza dyskryminacyjna*, moduł *Ogólne modele analizy dyskryminacyjnej* umożliwia również stosowanie analizy metodą najlepszego podzbioru. W szczególności można wybrać dobór predyktorów lub zbiorów predyktorów (w przypadku efektów o wielu stopniach swobody, uwzględniających predyktory jakościowe) metodą krokową i metodą najlepszego podzbioru w oparciu o statystykę  $F$  do wprowadzania i  $p$  do wprowadzania (powiązaną ze statystyką wielowymiarowego testu Lambda Wilksa). Ponadto, jeśli określimy próbę do oceny krzyżowej, wówczas dobór metodą najlepszego podzbioru możemy przeprowadzić w oparciu o wskaźniki błędnych klasyfikacji dla tej próby. Oznacza to, że po oszacowaniu funkcji dyskryminacyjnych dla danego zbioru predyktorów obliczane są wskaźniki błędnych klasyfikacji i wybierany jest ten model (podzbiór predyktorów), który daje najniższą wartość wskaźnika błędnych

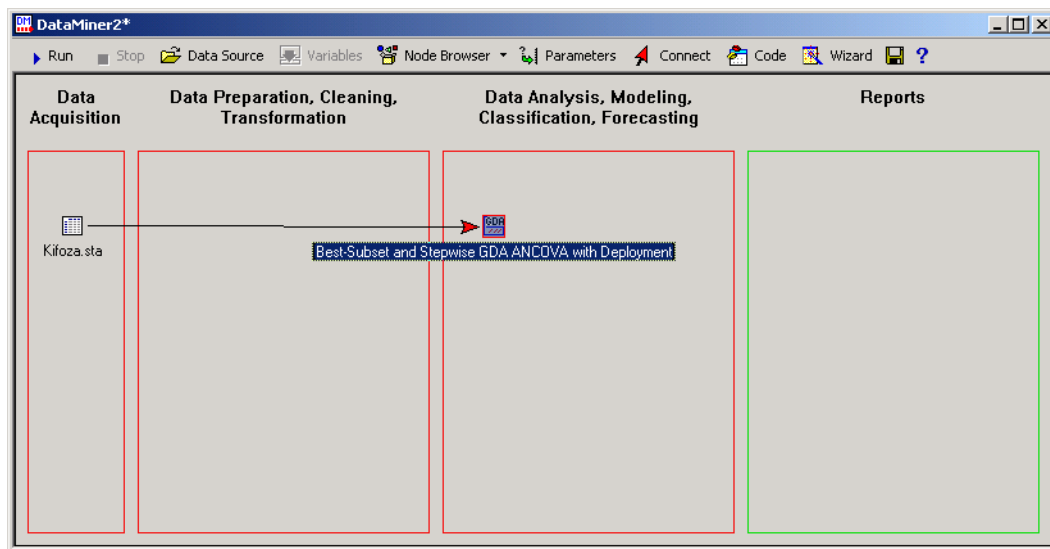


klasyfikacji w próbie przeznaczony do oceny krzyżowej. Jest to zatem potężna technika umożliwiająca wybór modeli charakteryzujących się dobrą trafnością prognostyczną i pozwalająca jednocześnie uniknąć nadmiernego dopasowania modelu do danych.

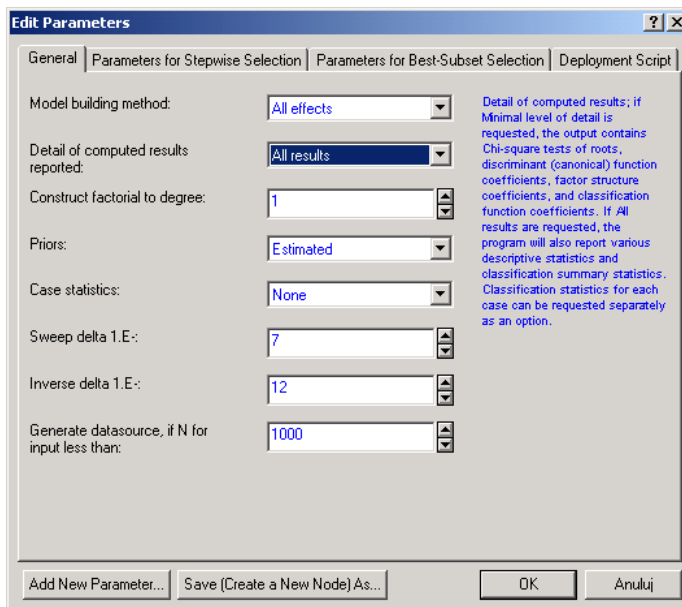
Moduł *Ogólne modele analizy dyskryminacyjnej* zawiera opcje, które czynią z tej techniki niezwykle efektywne narzędzie do zagadnień klasyfikacyjnych i technik *zglębiania danych* (*data mining*). Wykorzystanie metody najlepszego podzbioru (w szczególności w powiązaniu z predyktorami jakościowymi lub w przypadku wskaźników błędnych klasyfikacji w próbie przeznaczony do oceny krzyżowej) do wyboru najlepszego podzbioru predyktorów powinno być traktowane bardziej jako metoda heurystycznego poszukiwania niż klasyczna technika statystyczna.

Tak jak to zostało wcześniej zapowiedziane, przy budowie modelu klasyfikacyjnego użyjemy programu *STATISTICA Data Miner*.

Rozpoczynając budowanie projektu data mining, w panelu *Źródło danych* wskazujemy plik danych o nazwie *Kifoza*. 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 jakościowej użyjemy zmiennej o nazwie **Kifoza**, natomiast jako predyktory ciągłe wskazujemy zmienne: **Wiek**, **Początkowy krąg** oraz **Liczba kręgów**. Kliknięciem przycisku *OK* akceptujemy wybory, a następnie w oknie *Select dependent variables and predictors* jeszcze raz klikamy przycisk *OK*. Spowoduje to powrót do obszaru projektów. W drugim etapie budowania projektu moglibyśmy, gdyby zaistniała taka potrzeba, zastosować procedury czyszczące lub przekształcające surowe dane. W naszym przypadku nie ma takiej potrzeby. W kolejnym etapie wybieramy rodzaj analizy, którą chcemy zastosować w tworzonym projekcie. W tym celu korzystamy z przycisku *Node Browser (Przeglądarka węzłów)*. Przeglądarka węzłów swoją funkcjonalnością przypomina Eksplorator Windows. Na ekranie pojawi się okno o tej samej nazwie, w którym możemy wybrać konkretną analizę. Następnie z rozwijanej listy umieszczonej bezpośrednio pod paskiem tytułowym opisywanego okna wybieramy pozycję *All proceduras*. Aby wybrać odpowiedni rodzaj analizy, „wchodzimy” do folderu *Classification and Discrimination*. W prawym panelu zaznaczamy następnie pozycję *Best-Subset and Stepwise GDA ANCOVA with deployment*. Aby wstawić wybraną analizę do obszaru roboczego projektu, klikamy przycisk *Insert into workspace*. Po zamknięciu tego okna powracamy do obszaru roboczego. Ikonka symbolizująca źródło danych została połączona strzałką z ikonką symbolizująca wybrany przez nas rodzaj analizy (jak to pokazano na rysunku poniżej).



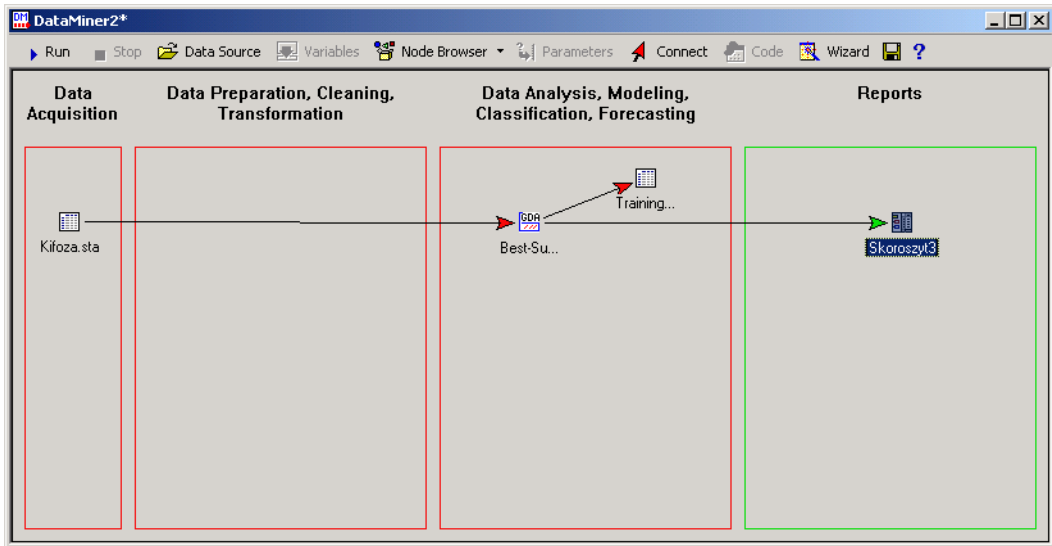
Aby obejrzeć lub zmienić parametry związane z wybraną analizą, klikamy dwukrotnie lewym przyciskiem myszy ikonkę symbolizującą analizę. Na ekranie pojawi się pokazane poniżej okno *Edit parameters*.



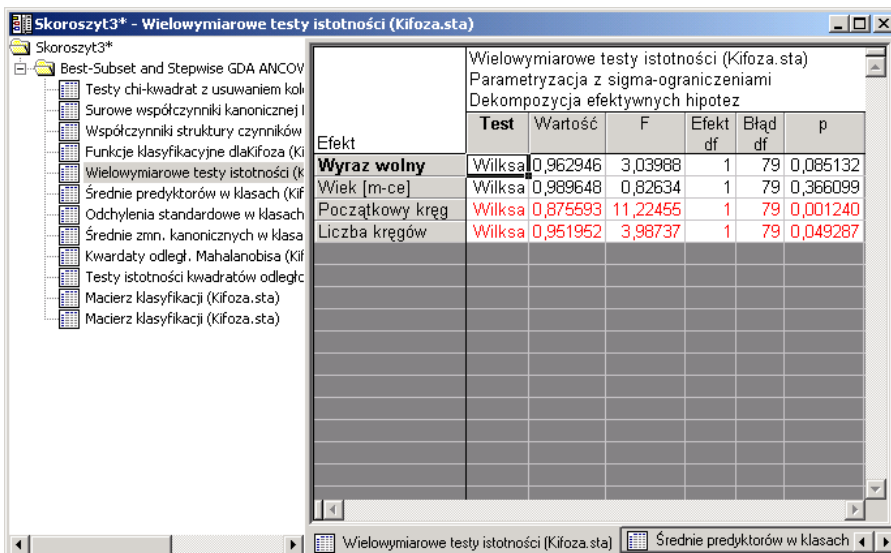
W oknie tym rozwijamy listę *Detail of computed results reported* i wybieramy opcję *All results*. Umożliwia to otrzymanie szerszego zakresu wyników analizy. Po zaakceptowaniu zmian (kliknięciem przycisku *OK*) wracamy znów do obszaru roboczego. W tym momencie projekt analizy jest gotowy. W celu wykonania zaprojektowanej analizy i uzyskania wyników klikamy przycisk *Run*. Po przeprowadzeniu analizy w panelu *Reports* program



umieszcza skoroszyt, zawierający wszystkie żądane wyniki. Poniżej przedstawiono wygląd obszaru roboczego po przeprowadzeniu analizy.



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.



Przystępując do oglądania wyników analizy, zaczniemy od tabeli, w której są podawane funkcje klasyfikacyjne. Są one wykorzystywane do konstruowania przewidywanej klasyfikacji.



Dane: Funkcje klasyfikacyjne dla kifoza (Kifoza)*		
Funkcje klasyfikacyjne dla kifoza (Kifoza.sta)		
Parametryzacja z sigma-ograniczeniami		
Efekt	występuje p=,2169	nie występuje p=,7831
<b>Wyraz wolny</b>	-11,4407	-10,3690
Wiek [m-ce]	0,0247	0,0200
Początkowy krąg	0,6610	0,8997
Liczba kręgów	2,2766	1,9217

Następnie obejrzymy wyniki testowania istotności ocen parametrów otrzymanego modelu.

Dane: Wielowymiarowe testy istotności (Kifoza)*						
Wielowymiarowe testy istotności (Kifoza.sta)						
Parametryzacja z sigma-ograniczeniami						
Dekompozycja efektywnych hipotez						
Efekt	Test	Wartość	F	Efekt df	Błąd df	p
<b>Wyraz wolny</b>	Wilksa	0,962946	3,03988	1	79	0,085132
Wiek [m-ce]	Wilksa	0,989648	0,82634	1	79	0,366099
Początkowy krąg	Wilksa	0,875593	11,22455	1	79	0,001240
Liczba kręgów	Wilksa	0,951952	3,98737	1	79	0,049287

Wyniki zamieszczone w powyższej tabeli sugerują, że z modelu należy usunąć zmienną **Wiek**, gdyż odpowiadający jej poziom istotności przekracza domyślną wartość 0,05. Aby tego dokonać, zmodyfikujemy parametry analizy. W obszarze roboczym klikamy dwukrotnie ikonkę symbolizującą analizę, a następnie na rozwijanej liście *Model building method* wybieramy pozycję *Forward stepwise* i ponownie uruchamiamy analizę.

Dane: Wielowymiarowe testy istotności (Kifoza)*						
Wielowymiarowe testy istotności (Kifoza.sta)						
Parametryzacja z sigma-ograniczeniami						
Dekompozycja efektywnych hipotez						
Efekt	Test	Wartość	F	Efekt df	Błąd df	p
<b>Wyraz wolny</b>	Wilksa	0,948148	4,37498	1	80	0,039642
Wiek [m-ce]	Wilksa	1,000000		0		
Początkowy krąg	Wilksa	0,879270	10,98456	1	80	0,001382
Liczba kręgów	Wilksa	0,952424	3,99618	1	80	0,048999

Oglądając ponownie wyniki analizy (patrz tabela powyżej), widzimy, że wszystkie oceny parametrów modelu istotnie różnią się od zera. Możemy teraz ocenić prognozowaną klasyfikację.

Dane: Macierz klasyfikacji (Kifoza)*			
Macierz klasyfikacji (Kifoza.sta)			
Wiersze: obserwowana klasyfik.			
Kolumny: przewidywana klasyfikacja			
Klasa	Procent Poprawne	występuje p=,2169	nie występuje p=,7831
<b>występuje</b>	38,88889	7,00000	11,00000
nie występuje	90,76923	6,00000	59,00000
Ogół	79,51807	13,00000	70,00000

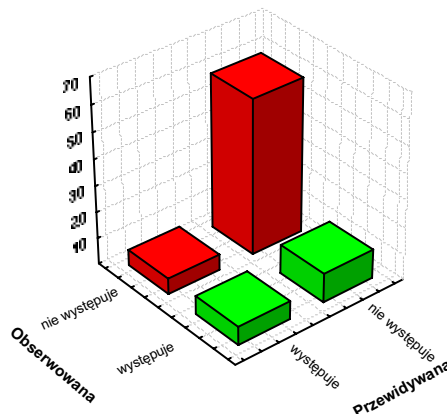
W boczkuz zamieszczonej powyżej tabeli widać obserwowaną klasyfikację pacjentów, natomiast w nagłówku tabeli klasyfikację przewidywaną. Tak więc uzyskany model pozwala poprawnie sklasyfikować 7 przypadków (co stanowi 38,9% wszystkich tych przypadków, u których występowała kifoza) spośród tych, u których rzeczywiście występowała kifoza, oraz błędnie przewiduje niewystępowanie kifozy u 11 przypadków (61,1%). Jednocześnie model pozwala poprawnie sklasyfikować 59 przypadków (90,8%) spośród tych, u których kifoza rzeczywiście nie występowała, oraz błędnie klasyfikuje występowanie kifozy w przypadku 6 przypadków (9,2%).

Wyniki analizy są umieszczane w standardowych arkuszach programu *STATISTICA*. Na tych wynikach możemy wykonywać dalsze analizy. Aby to zilustrować przedstawimy uzyskaną klasyfikację za pomocą dwuwymiarowego wykresu słupkowego. W tym celu w arkuszu przedstawiającym macierz klasyfikacji zaznaczamy komórki zawierające odpowiednie liczebności (przedstawia to poniższy zrzut ekranu).

Klasa	Procent	występuje	nie występuje
	Poprawne	p=,2169	p=,7831
występuje	38,88889	7,00000	11,00000
nie występuje	90,76923	6,00000	59,00000
Ogół	79,51807	13,00000	70,00000

W kolejnym kroku klikamy prawym przyciskiem myszy wewnątrz zaznaczonego obszaru i z podręcznego menu wybieramy kolejno opcje *Wykresy bloku danych* oraz *Wykres użytkownika z bloku, wierszami*. Na ekranie pojawi się okno *Wybierz wykres*. W oknie tym jako kategorię wykresu wybieramy *Wykresy sekwencyjne*, jako rodzaj wykresu *Wykresy surowych danych*, a jako podtyp wykresu *Kolumny*. Otrzymany w efekcie wykres przedstawiono poniżej.

Wykres klasyfikacji obserwowanej i przewidywanej





Przedstawia on w sposób graficzny opisaną wcześniej klasyfikację przypadków.

W kolejnej części zajmiemy się zbudowaniem modelu klasyfikacyjnego za pomocą techniki drzew klasyfikacyjnych.

## Budowa modelu klasyfikacyjnego przy użyciu drzew klasyfikacyjnych

Drzewa klasyfikacyjne wykorzystuje się do wyznaczania przynależności przypadków lub obiektów do klas jakościowej zmiennej zależnej na podstawie wartości jednej lub większej liczby zmiennych objaśniających (predyktorów). Analiza drzew klasyfikacyjnych jest jedną z podstawowych technik wykorzystywanych w tzw. zgłębianiu danych (data mining). Moduł *Drzew klasyfikacyjnych* zawarty w programie *STATISTICA* jest kompletną implementacją technik obliczania binarnych drzew klasyfikacyjnych w oparciu o podziały jednowymiarowe dla predyktorów nominalnych, predyktorów porządkowych (mierzonych przynajmniej na skali porządkowej) lub obu typów predyktorów łącznie.

Celem analizy opartej na drzewach klasyfikacyjnych jest przewidywanie lub wyjaśnianie odpowiedzi (reakcji) zakodowanych w jakościowej zmiennej zależnej i dlatego techniki wykorzystywane w tym module mają wiele wspólnego z technikami wykorzystywanymi w bardziej tradycyjnych metodach, takich jak np. analiza dyskryminacyjna czy też analiza skupień. Elastyczność analizy przeprowadzanej za pomocy drzew klasyfikacyjnych sprawia, że jest ona bardzo atrakcyjna, ale nie oznacza to, że zaleca się stosowanie jej zamiast metod bardziej tradycyjnych. Jeśli surowsze wymogi teoretyczne i założenia dotyczące rozkładów, wymagane przez metody tradycyjne, są spełnione, lepiej wykorzystywać te metody. Jednakże drzewa klasyfikacyjne stanowią niezrównaną technikę eksploracyjną i, gdy zawiodą metody tradycyjne, bywają ostatnią „deską ratunku”.

Co to są drzewa klasyfikacyjne? Wyobraźmy sobie, że chcemy wypracować system sortowania zestawu monet w różne klasy (dziesięciogroszówki, dwudziestogroszówki, pięćdziesięciogroszówki i złotówki). Załóżmy, że istnieje miara, którą różnią się te monety, na przykład średnica, którą możemy wykorzystać do opracowania hierarchicznego systemu sortowania monet. Moglibyśmy potoczyć monety w dół wąskim torem, w którym wycięto otwór wielkości średnicy dziesięciogroszówki. Jeśli moneta wypadnie przez otwór, zostanie zaklasyfikowana jako dziesięciogroszówka, w przeciwnym razie potoczy się dalej, gdzie znajduje się otwór o wielkości średnicy dwudziestogroszówki. Jeśli moneta wypadnie przez ten otwór, zostanie zaklasyfikowana jako dwudziestogroszówka, jeśli nie, potoczy się dalej, gdzie znajduje się otwór o wielkości pięćdziesięciogroszówki i tak dalej. W ten sposób zbudowaliśmy drzewo klasyfikacyjne. Proces decyzyjny wykorzystany w naszym drzewie klasyfikacyjnym stanowi skuteczną metodę sortowania stosu monet, a ogólniej, może być zastosowany do rozwiązania wielu problemów klasyfikacyjnych.

Analiza i wykorzystanie drzew klasyfikacyjnych nie są szeroko rozpowszechnione w teorii prawdopodobieństwa i statystycznym rozpoznawaniu obrazów, ale drzewa klasyfikacyjne są szeroko wykorzystywane w wielu dziedzinach tak odmiennych jak: medycyna (diagnoza), nauki komputerowe (struktury danych), botanika (klasyfikacja) i psychologia



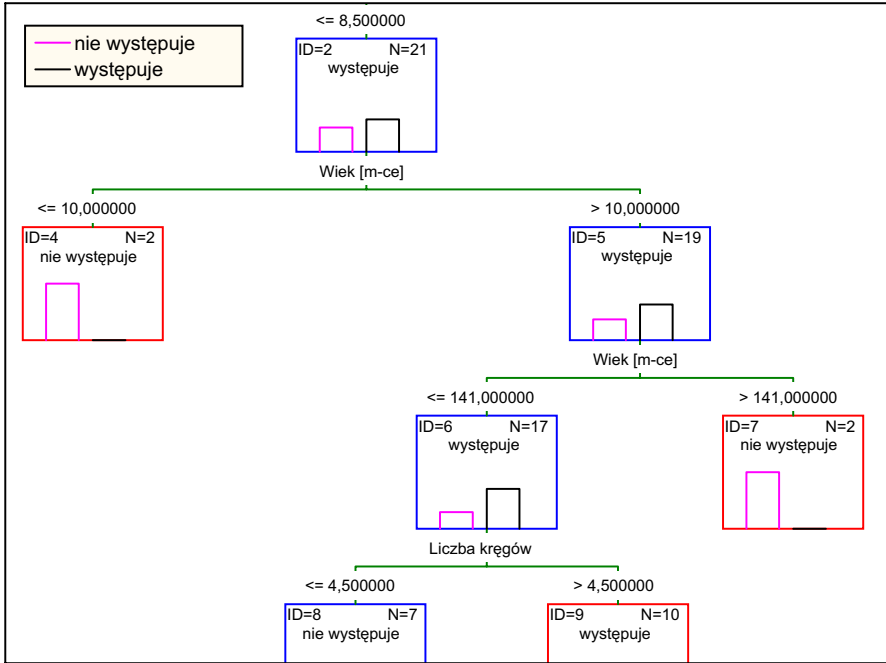
(teoria decyzji). Drzewa klasyfikacyjne dają się prosto przedstawiać graficznie, co sprawia, że są łatwiejsze w interpretacji niż wyniki czysto liczbowe.

Podobnie jak w poprzednim przykładzie, przy budowie odpowiedniego modelu klasyfikacyjnego wykorzystamy środowisko programu *STATISTICA Data Miner*. Jak poprzednio, rozpoczynając budowanie projektu data mining, w panelu *Źródło danych* wskazujemy plik danych o nazwie *Kifoza*. Następnie w oknie *Select dependent variables and predictors*, które pojawi się na ekranie, wskazujemy zmienną zależną i predyktory (zmienne objaśniające) w taki sam sposób, jak w przypadku budowy modelu za pomocą ogólnej analizy dyskryminacyjnej. W kolejnym etapie wybieramy rodzaj analizy, którą chcemy wykorzystać w tworzonym projekcie. Korzystając z *Przeglądarki węzłów*, zaznaczamy pozycję *Standard Classification Trees with Deployment (C and RT)*. Aby wstawić wybraną analizę do obszaru roboczego projektu, korzystamy z przycisku *Insert into workspace*. Po zamknięciu tego okna powracamy do obszaru roboczego.

Aby obejrzeć lub zmienić parametry związane z wybraną analizą, klikamy dwukrotnie lewym przyciskiem myszy ikonkę symbolizującą analizę. Na ekranie pojawi się okno *Edit parameters*. W oknie tym rozwijamy listę *Detail of computed results reported* i wybieramy opcję *Comprehensive*. Umożliwia to otrzymanie szerszego zakresu wyników analizy. Po zaakceptowaniu zmian (kliknięciem przycisku *OK*) wracamy znów do obszaru roboczego. W tym momencie projekt analizy jest gotowy. W celu wykonania zaprojektowanej analizy i uzyskania wyników klikamy przycisk *Run*. Po przeprowadzeniu analizy w panelu *Reports* program umieszcza skoroszyt, zawierający wszystkie żądane wyniki.

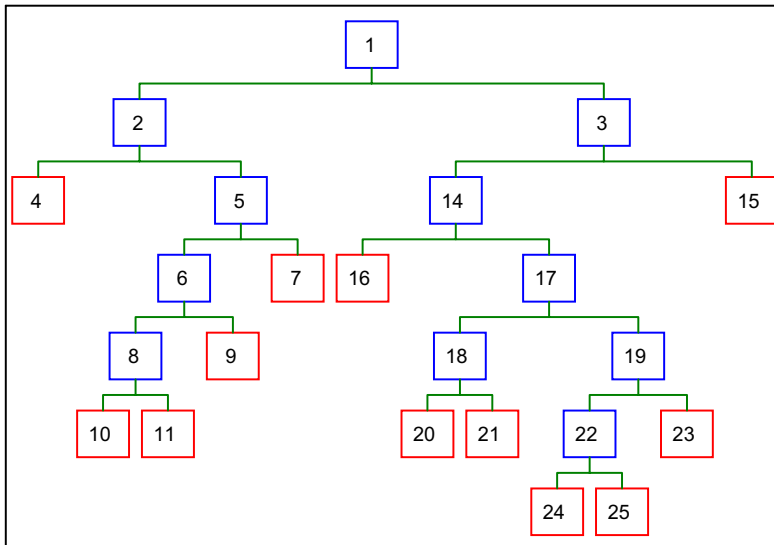
W celu obejrzenia otrzymanych wyników klikamy dwukrotnie ikonkę symbolizującą skoroszyt z wynikami analizy. Na ekranie pojawi się okno skoroszytu. Zaznaczając odpowiednią pozycję w lewym panelu skoroszytu, możemy przeglądać odpowiednie wyniki.

Zacznijmy od oglądnięcia drzewa klasyfikacyjnego. Przedstawia ono wynik rekurencyjnego podziału całego zbioru obiektów na rozłączne podzbiory, a jego hierarchiczna struktura najlepiej oddaje sekwencje kolejnych kroków procedury podziału. Otrzymane drzewo składa się z 24 wierzchołków, z których 12 to tzw. liście, czyli węzły końcowe. Do dokładnego przeglądania drzewa dostępne są specjalne narzędzia, za pomocą których można np. powiększyć wybrany fragment drzewa. Na rysunku poniżej został zamieszczony taki przykładowy fragment.



Możemy też zobaczyć ogólną strukturę drzewa. Odpowiedni wykres przedstawia w postaci schematu wszystkie wierzchołki.

Struktura drzewa dla pliku Kifoza



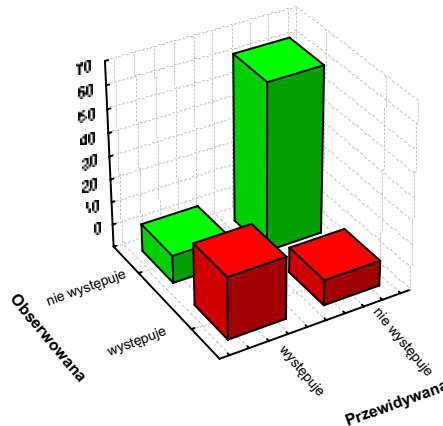
Możemy również ocenić wyniki klasyfikacji otrzymywanej w wyniku zastosowania oszacowanego modelu. Umożliwia to macierz klasyfikacji przypadków. Zawiera ją zamieszczony poniżej arkusz.

		Class	
		nie występuje	występuje
nie występuje	nie występuje	63,00000	1,00000
	występuje	2,00000	17,00000

Wyniki zamieszczone w arkuszu pokazują otrzymaną klasyfikację. W boczku zamieszczonej powyżej tabeli widać obserwowaną klasyfikację pacjentów, natomiast w nagłówku tabeli klasyfikację przewidywaną. Tak więc uzyskany model pozwala poprawnie sklasyfikować 17 przypadków spośród tych, u których rzeczywiście występowała kifoza, oraz błędnie przewiduje niewystępowanie kifozy w 2 przypadkach. Jednocześnie model pozwala poprawnie sklasyfikować 63 przypadki spośród tych, u których kifoza rzeczywiście nie występowała, oraz błędnie klasyfikuje występowanie kifozy w 1 przypadku. Tak więc uzyskane wyniki wskazują, że model drzew klasyfikacyjnych pozwala na poprawniejszą klasyfikację przypadków w porównaniu z modelem uzyskanym w wyniku zastosowania ogólnej analizy dyskryminacyjnej.

Poniżej zamieszczono również wykres prezentujący w sposób graficzny opisaną wcześniej klasyfikację przypadków.

Wykres klasyfikacji obserwowanej i przewidywanej



Na końcu warto jeszcze raz przypomnieć, że podobne modele można budować także dla danych pochodzących z innych dziedzin. Mogłyby to być np. modele ujmujące czynniki wpływające na zapadnięcie na dana chorobę w medycynie, czynniki decydujące o osiągnięciach szkolnych w zagadnieniach pedagogicznych, czynniki wpływające na spłatę kredytu



w zagadnieniach bankowych czy też czynniki wpływające na powstanie szkody w zagadnieniach ubezpieczeniowych.

## **Bibliografia**

1. Berry M. J. A., Linoff G. S., 2000, *Mastering Data mining. The Art and Science of Customer Relationship Management*, Wiley.
2. Gatnar E., 2001, *Nieparametryczna metoda dyskryminacji i regresji*, PWN Warszawa.
3. Hastie T. J., Tibshirani R. J., 1990, *Generalized Additive Models*, Chapman & Hall/CRC.
4. Hastie T. J., Tibshirani R. J., 2001, *The Elements of Statistical Learning*, Springer.
5. *STATISTICA Data Miner Manual*, StatSoft, Inc., 2002.