Prognozowanie kondycji ekonomiczno-finansowej przedsiębiorstw z wykorzystaniem drzew decyzyjnych

A. Bożek[1]

1. Wstęp.

Identyfikacja czynników mających wpływ na powodzenie bądź niepowodzenie działalności przedsiębiorcy jest jedną z najważniejszych umiejętności potrzebnych w każdej firmie. Wielość czynników oddziaływujących na poszczególne przedsiębiorstwa powoduje, że decydentom trudno o ich prawidłową analizę i ocenę a przede wszystkim prawidłowe określenie ich wpływu na zarządzaną przez nich jednostkę. Wykorzystanie najnowszych osiągnięć nauki może być pomocne w identyfikacji przesłanek determinujących przyszłą kondycję przedsiębiorstwa.

2. Uczenie maszynowe przy użyciu drzew decyzyjnych.

Jednym z naukowych podejść, które mogą być użyteczne przedsiębiorstwom w prognozowaniu przyszłej kondycji finansowej są metody i narzędzia Machine Learning. Uczenie maszynowe (ML) to jedna z najważniejszych poddziedzin sztucznej inteligencji, która łączy rozwiązania z dziedziny statystyki, informatyki, nauk kognitywnych, teorii rozpoznawania i wielu innych dziedzin [3] . Rozwinięte w latach dziewięćdziesiątych minionego stulecia metody Data Mining określane w Polsce metodami eksploracji danych (drążenia danych, odkrywania zależności w bazach danych) to jedne z najszerzej stosowanych narzędzi informatycznych (oczywiście poza narzędziami służącymi do administrowania, gromadzenia i przekazywania informacji) w obecnym czasie. Metody te są zawarte w nowoczesnych aplikacjach i służą średniemu i najwyższemu szczeblowi zarządzania do podejmowania decyzji w oparciu o wiedzę „wyszukaną" z wewnętrznej dokumentacji organizacji oraz wyników przeprowadzonych badań.

Zastosowanie metod uczenia maszynowego sprowadza się do trzech kroków [13]:

- przygotowania danych (zbiór uczący, zbiór testowy),
- analizy danych (budowa modelu)
- wdrożenia.

Pierwszym etapem procesu eksploracji jest przygotowanie danych, czyli czyszczenie i przekształcanie, wybór podzbiorów rekordów (przypadków), ewentualny wstępny wybór zmiennych (cech), którego celem jest inteligentne zredukowanie wielkości danych. Do eksploracji wykorzystywany jest szeroki wachlarz metod, od regresji liniowej do zaawansowanych metod statystycznych. Efektem tego etapu są przygotowane do analizy dwa zbiory: uczący i testowy, a przy wykorzystaniu bardziej złożonych narzędzi dataminingowych również zbiór walidacyjny.

Drugi etap projektu data mining to budowanie modelu i jego ocena. Buduje się tu różne modele, wybierając najlepszy z nich, czyli taki w którym błąd dopasowania modelu jest jak najmniejszy. Najczęściej stosowane techniki to: agregacja modeli, czyli głosowanie i uśrednianie, wzmacnianie, kontaminacja modeli i metauczenie (meta-learning). Do budowy modeli wykorzystywane są różne algorytmy od ID3 do CHAID lub nowszych.

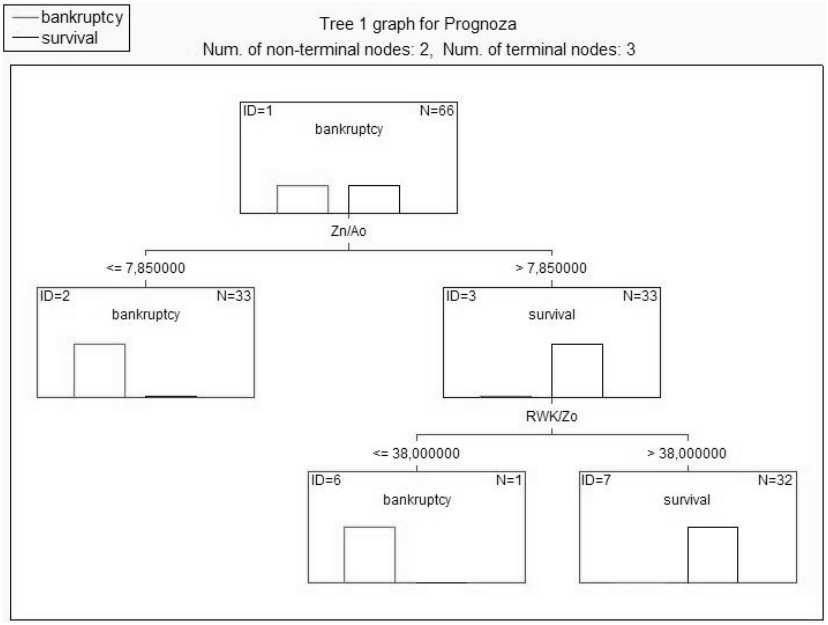
Trzeci i ostatni etap to wdrożenie modelu, czyli zastosowanie do nowych przypadków w celu oceny nowych danych według kryteriów stosowanych w modelu. Najnowsze aplikacje, np. Statistica Data Miner generują kod C++, który może być wykorzystany do budowy miniaplikacji (kalkulatorów) wyliczających aktualne wartości modelu.

Jedną z metod konstruowania modeli machine learning jest metoda drzew decyzyjnych. Jest to jedna z najpopularniejszych i najbardziej skutecznych metod drążenia danych, która bardzo często stosowana jest do predykcji. Drzewa są wykorzystywane do określania przynależności obiektów do klas na podstawie pomiarów jednej lub wielu zmiennych opisujących, określając ich wpływ na jakościową zmienną zależną – zmienną prognozowaną (przewidywaną).

Drzewa klasyfikacyjne tworzone są wtedy, gdy zmienna zależna ma charakter jakościowy a drzewa regresyjne – przy ciągłej postaci zmiennej zależnej. Tą metodą uczenia maszynowego poszukiwane są takie części przestrzeni cech parametrów, w których zmienna zależna przyjmuje tylko pewną określoną wartość (zmienna jakościowa np.: dobra lub zła kondycja przedsiębiorstwa). Drzewa klasyfikacyjne poszukują podobnych reguł, z tym że są w stanie znaleźć je w bardzo skomplikowanych wielowymiarowych przypadkach w przeciwieństwie do zdolności człowieka w naocznym wyszukiwaniu zależności. Otrzymane reguły standardowo prezentuje się w postaci drzewa, dzięki czemu są one przejrzyste, nawet w przypadku rozległych drzew.

Korzeń powyższego drzewa (umieszczony na górze, gdyż drzewo jest odwrócone) reprezentowany jest liczbą wszystkich badanych rekordów, wierzchołki wewnętrzne określają sposób dokonywania podziału w oparciu o wartości cech obiektów. Liście (węzły zewnętrzne) reprezentują klasy, do których należą obiekty. Krawędzie drzewa wskazują wartości cech, na podstawie których dokonywany jest podział. W powyższym drzewie zawarte są informacje pozwalające odczytać reguły przynależności rekordów do klas. Widać, że najważniejszą zmienną predykcyjną jest rentowność majątku. Rozdziela ona zbiór wszystkich rekordów wartością 7,85. Jedna z reguł decyzyjnych tej analizy jest następująca: jeżeli wskaźnik [Zn/Ao] jest większy od 7,85 i wskaźnik

[1] Wyższa Szkoła Zarządzania i Administracji w Zamościu, Katedra Nauk Ekonomicznych.



Rys. 1. Przykładowe drzewo decyzyjne

[RWK/Zo] jest większy od 38; firma nie zbankrutuje, istnieje 32 takich obserwacji.

Najważniejsze zalety drzew decyzyjnych to:

- szybkość analizy (czas decyzyjny ograniczony liniowo liczbą atrybutów),
- prosta forma reprezentacji reguł (łatwość zrozumienia wyników),
- odkrycie łatwych w interpretacji reguł,
- łatwość stosowania algorytmu ze zrozumieniem nawet dla osób bez dużego doświadczenia w analizie danych,
- mogą reprezentować dowolnie złożone pojęcia pojedyncze lub wielokrotne.
- odporność na nawet dużą liczbę predyktorów nie mających wpływu na badaną zmienną

Podstawową wadą drzew decyzyjnych jest fakt, że testuje się za każdym razem wartość tylko jednego atrybutu, co powoduje niepotrzebny rozrost drzewa dla danych gdzie poszczególne atrybuty zależą od siebie.

Tworzenie drzewa polega na rekurencyjnym podziale zbioru uczącego (zawierającego obiekty o których wiadomo jest do jakich klas należą) na podzbiory aż do uzyskania ich jednorodności ze względu na przynależność obiektów do klas. Celem jest tu utworzenie drzewa o jak najmniejszej liczbie węzłów, aby otrzymać jak najprostsze reguły klasyfikacyjne.

Drzewa klasyfikacyjne i regresyjne poszukują optymalnego podziału na segmenty, stosując poniższy schemat działania [8]:

1. W zbiorze obiektów S, sprawdzenie, czy należą do tej samej klasy. Jeżeli tak, zakończenie postępowania.
2. Jeśli nie, rozważenie wszystkich możliwych podziałów zbioru S na rozłączne podzbiory S_1, S_2, ... S_s tak, by były jak najbardziej jednorodne (s – liczba podzbiorów)
3. Ocena jakości każdego z tych podziałów zgodnie z przyjętym kryterium i wybór najlepszego z nich;
4. Podział zbiór S w wybrany powyżej sposób;
5. Wykonanie kroków 1-4 rekurencyjnie, przyjmując jako S każdy z otrzymanych podzbiorów S_1, S_2, ... S_s.

Jako reguły stopu stosuje się m.in.: minimalną liczność węzła podlegającego podziałom, minimalną liczność węzła powstającego w wyniku podziałów i maksymalną liczbę poziomów drzewa. Po zakończeniu podziałów wykonuje się jeszcze operację doboru właściwej wielkości drzewa, np. przycinanie (pruning). Przycinanie polega na usuwaniu gałęzi drzewa, co wykonujemy automatycznie lub ręcznie, w oparciu o posiadaną wiedzę o celach analizy, jakości pomiaru poszczególnych cech, ograniczeniach stosowania modelu itp. (jest to wiedza, której nie ma w danych i siłą rzeczy analiza danych nie może jej wydobyć). Końcowym efektem takiej analizy powinno być utworzenie drzewa o możliwie najmniejszej liczbie gałęzi i węzłów aby znaleźć możliwie najprostsze reguły klasyfikacyjne.

Głównym kryterium podziału przestrzeni cech jest funkcja oceniająca jakość podziału (stopień jednorodności podzbiorów), która jest maksymalizowana. Alternatywnie, algorytm może szukać minimum funkcji mierzącej niejednorodność (misclasification).

Stosowanymi algorytmami podziału (wg heurystycznego schematu TDIDT – Top Down Induction of Decision Tree) są [6] :

- ID3 (Quinlan 1983) wykorzystywany z modyfikacjami w DTReg
- C4.8 (Quinlan 1986)
- C&RT (system CART)
- Assistant

- QUEST
- CHAID (Chi-squared Automatic Interaction Detection, Biggs, DeVille, Suen 1991) - system AnswerTree, SPSS, Statistica Data Miner

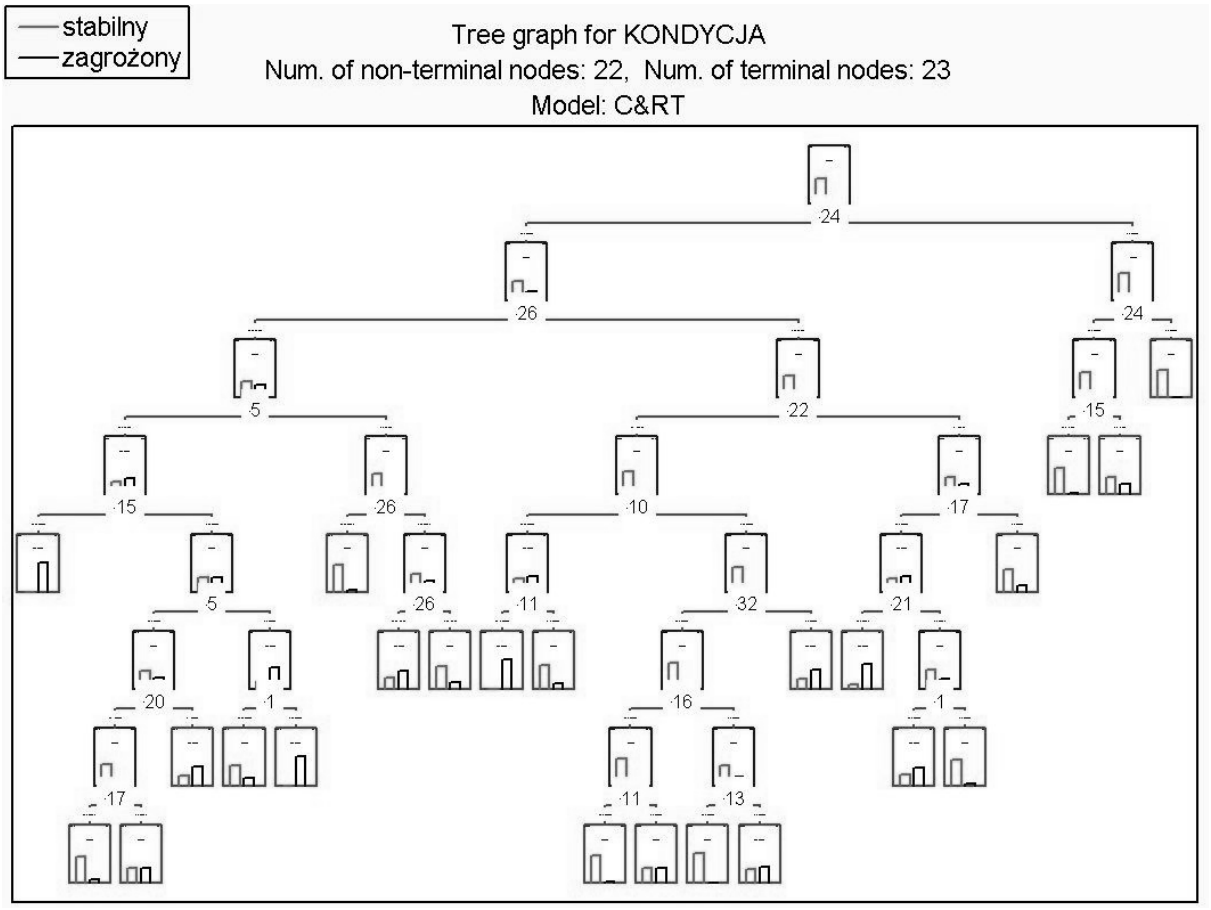
Różnice między konkretnymi algorytmami dotyczą przede wszystkim sposobu wyboru testu dla węzła związanego z oceną jakości podziału zbioru przykładów w węźle, zasad podejmowania decyzji o utworzeniu liścia lub węzła oraz technik uwzględniania różnego rodzaju zaburzeń w opisie przykładów uczących.

Jako, że proste drzewa klasyfikacyjne i regresyjne (generujące jedno drzewo decyzyjne) nie są niekiedy w stanie opisać wszystkich złożonych zależności, metodę drzew można wzbogacić innymi, bardziej skomplikowanymi procedurami. Do wykorzystywanych najczęściej należą [7]: ważenie, wzmacnianie (boosting), zespoły drzew decyzyjnych (decision tree forest), metoda wektorów nośnych (support vector machine - SVM), regresja logistyczna, V-krotny sprawdzian krzyżowy oraz metoda globalnego sprawdzianu krzyżowego.

Poszukiwanie i testowanie modelu skupia się na minimalizowaniu błędnych podziałów węzłów w drzewach decyzyjnych. Tak więc wykorzystywane są te metody, które dla zadanego zestawu uczącego i testowego cechują się najlepszym dopasowaniem wyników modelu do rzeczywistych klas obiektów.

3. Zastosowanie metody drzew decyzyjnych do prognozy kondycji przedsiębiorstw handlowych województwa podkarpackiego.

Zastosowanie metody drzew decyzyjnych do prognozowania kondycji ekonomiczno–finansowej małych przedsiębiorstw zostanie zaprezentowane na przykładzie prognozy kondycji firm sklasyfikowanych w sekcji G – handel hurtowy i detaliczny z terenu województwa podkarpackiego. W prezentowanym badaniu przeanalizowano 1893 przypadków małych przedsiębiorstw, które w latach 1999 – 2004 składały w dwóch kolejnych latach sprawozdania i nie były w tym czasie w stanie likwidacji lub upadłości. Do badań został wykorzystany system Statistica Data Miner, który jest zestawem narzędzi w formie modułów drążenia danych zaimplementowanych do jednego z najlepszych i najczęściej stosowanych systemów do analizy danych - Statistica. System umożliwia przygotowanie danych w postaci zbioru uczącego i testowego, intuicyjne prowadzenie przez procedurę budowy i dopasowywania modelu oraz przejrzystą wizualizację wyników badań. Statistica Data Miner oferuje również wdrożenie utworzonych modeli data mining w postaci kodu źródłowego w języku C++, SVB lub PMML.



Rys. 2. Drzewo klasyfikacyjne dla zmiennej KONDYCJA utworzone algorytmem C&RT

Pierwsze wyniki badań w oparciu o dane statystyczne w postaci drzewa decyzyjnego widoczne są na poniższym rysunku. Reguły decyzyjne, zaprezentowane są tutaj w widoku korzenia, gałęzi i liści. Pierwszy podział całej grupy obserwacji został przeprowadzony wartością wskaźnika rentowności sprzedaży (zmienna X_{24}) i punkt podziału tego wskaźnika ustalony został na poziomie 0,0003. Podział ten wyróżnił dwa kolejne węzły: nr 2 rozdzielony następnie wartością -0,0819 wskaźnika rentowności majątku (zmienna X_{26}) i nr 3 rozdzielony następnie wartością 0,003 wskaźnika rentowności sprzedaży (zmienna X_{24}). Podziały te wynikają z pracy algorytmu poszukującego podziału na możliwie jednorodne podzbiory.

Na poniższym rysunku widać wartości zmiennych, które dzieliły przypadki na możliwie jednorodne grupy. Drzewo składa się z 45 węzłów; 22 wewnętrznych (generujących kolejne podziały) i 23 końcowych. Powyższy podział ukazuje reguły decyzyjne, które program Statistica Data Miner udostępnia w postaci kodu źródłowego prezentowanego poniżej.

Najważniejszymi miernikami świadczącymi o kondycji przedsiębiorstwa są zmienne: X_{24} wskaźnik rentowności sprzedaży, X_{26} - wskaźnik rentowności majątku, X_{20} – wskaźnik intelektualnej wartości dodanej (VAIC) oraz zmienna X_{16} – wskaźnik pokrycia zobowiązań odsetkowych. Stosunkowo niewielki, acz istotny jest wpływ wskaźników mezo- i makroekonomicznych.

Tab. 1. Struktura drzewa klasyfikacyjnego

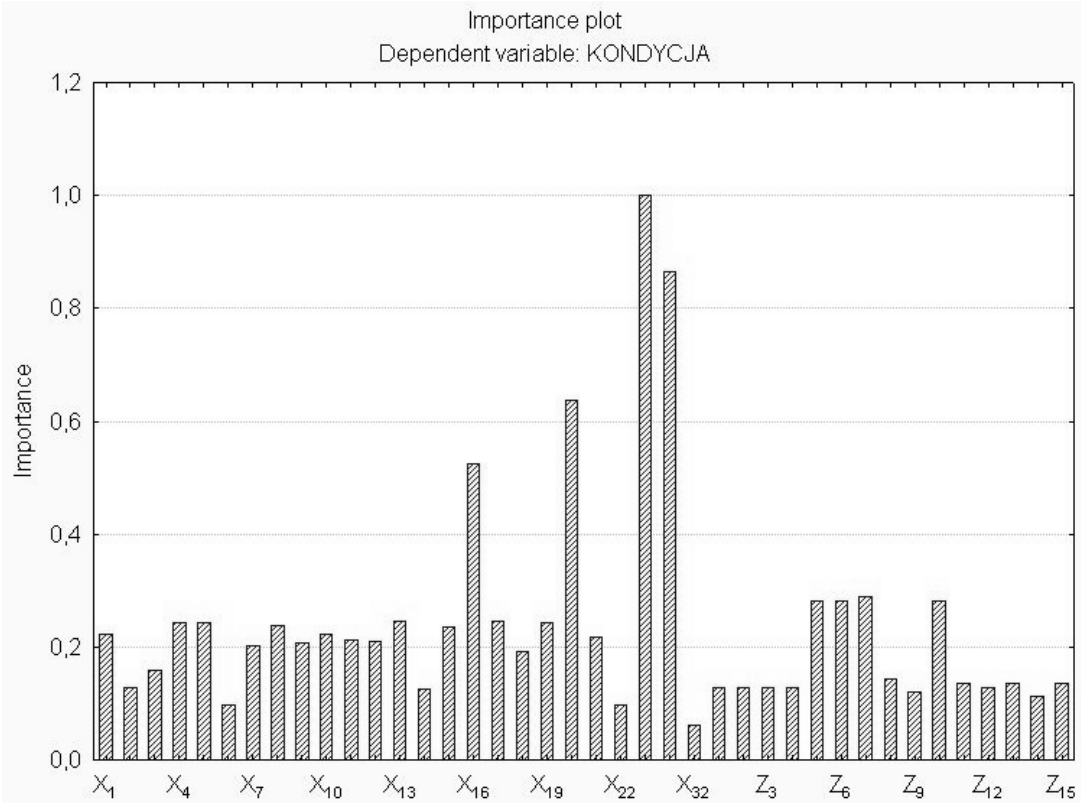
Tree Structure (Podział danych na próbę uczącą i testową (klasyfikacja))
Response: KONDYCJA
Model: C&RT

	Number of nodes	Size of node	N in class stabilny	N in class zagrożony	Selected category	Split variable	Criterion for child 1	Criterion for child 2	Child node 1	Child node 2
1	2	1415	1253	162	stabilny	X_{24}	x <= 0,0003	x > 0,0003	2	3
2	2	359	245	114	stabilny	X_{26}	x <= -0,0819	x > -0,0819	4	5
4	2	137	78	59	stabilny	X_{5}	x <= 3,8215	x > 3,8215	6	7
6	2	75	33	42	zagrożony	X_{15}	x <= -0,0604	x > -0,0604	8	9
8		10	0	10	zagrożony					
9	2	65	33	32	stabilny	Z_{5}	x <= 9,5000	x > 9,5000	10	11
10	2	45	28	17	stabilny	X_{20}	x <= 1,0550	x > 1,0550	12	13
12	2	33	24	9	stabilny	X_{17}	x <= 5,4824	x > 5,4824	14	15
14		19	17	2	stabilny					
15		14	7	7	stabilny					
13		12	4	8	zagrożony					
11	2	20	5	15	zagrożony	X_{1}	x <= 35,6205	x > 35,6205	16	17
16		7	5	2	stabilny					
17		13	0	13	zagrożony					
7	2	62	45	17	stabilny	X_{26}	x <= -0,2052	x > -0,2052	18	19
18		24	22	2	stabilny					
19	2	38	23	15	stabilny	X_{26}	x <= -0,1422	x > -0,1422	20	21
20		16	6	10	zagrożony					
21		22	17	5	stabilny					
5	2	222	167	55	stabilny	X_{22}	x <= 1,0581	x > 1,0581	22	23
22	2	154	125	29	stabilny	X_{10}	x <= -0,2696	x > -0,2696	24	25
24	2	19	9	10	zagrożony	Z_{11}	x <= 14,4000	x > 14,4000	26	27
26		8	0	8	zagrożony					
27		11	9	2	stabilny					
25	2	135	116	19	stabilny	X_{32}	x <= 0,2745	x > 0,2745	28	29
28	2	129	114	15	stabilny	X_{16}	x <= 0,0020	x > 0,0020	30	31
30	2	102	95	7	stabilny	X_{11}	x <= 0,0058	x > 0,0058	32	33
32		96	92	4	stabilny					
33		6	3	3	stabilny					
31	2	27	19	8	stabilny	X_{13}	x <= 0,4388	x > 0,4388	34	35
34		12	12	0	stabilny					
35		15	7	8	zagrożony					
29		6	2	4	zagrożony					
23	2	68	42	26	stabilny	X_{17}	x <= 2,2385	x > 2,2385	36	37
36	2	34	16	18	zagrożony	X_{21}	x <= 0,0067	x > 0,0067	38	39
38		13	2	11	zagrożony					
39	2	21	14	7	stabilny	Z_{1}	x <= 2,6000	x > 2,6000	40	41
40		10	4	6	zagrożony					
41		11	10	1	stabilny					
37		34	26	8	stabilny					
3	2	1056	1008	48	stabilny	X_{24}	x <= 0,0030	x > 0,0030	42	43
42	2	159	140	19	stabilny	X_{15}	x <= 0,1219	x > 0,1219	44	45
44		138	127	11	stabilny					
45		21	13	8	stabilny					
43		897	868	29	stabilny					

```
//     Continuous predictor name="Z14"; location=40
//     Continuous predictor name="Z15"; location=41
////////////////////////////////////////////////////////////////////////////////

      double ret;
      if(  Rnr[24] <= 3.25000000000000e-004 ) {
            if(  Rnr[25] <= -8.19500000000000e-002 ) {
                  if(  Rnr[6] <= 3.82147500000000e+000 ) {
                        if(  Rnr[16] <= -6.04350000000000e-002 ) {
                              ret = 1.02000000000000e+002;
                        }
                        else if(  Rnr[16] > -6.04350000000000e-002 ) {
                              if(  Rnr[31] <= 9.50000000000000e+000 ) {
                                    if(  Rnr[21] <= 1.05502500000000e+000 ) {
                                          if(  Rnr[18] <= 5.48241000000000e+000 ) {
                                                ret = 1.01000000000000e+002;
                                          }
                                          else if(  Rnr[18] > 5.48241000000000e+000
```

Rys. 3. Fragment kodu C++ umożliwiającego budowę aplikacji do oceniania nowych przypadków



Rys. 4. Istotność cech atrybutów

Błędna klasyfikacja zbioru testowego kształtuje się na poziomie 6,71%. Na 477 przypadków zbioru testowego błędnie sklasyfikowano 32 przypadki. Błędna klasyfikacja jest zazwyczaj niższa przy zastosowaniu wygenerowanych modeli do zbioru uczącego, wyższa przy zbiorze testowym.

Powyższe badania są kontynuowane w celu minimalizacji błędu złej klasyfikacji, minimalizacji ilości reguł decyzyjnych oraz wygenerowania optymalnego kodu źródłowego analizy dla kalkulatora kondycji finansowej przedsiębiorstw. Kalkulator ten umieszczony na stronie internetowej projektu, może zostać pomocnym źródłem informacji na temat kondycji danego przedsiębiorstwa.

Tab. 2. Błędna klasyfikacji dla zbioru uczącego - sekcja G

Classification matrix (Podział danych na próbę uczącą i testową (klasyfikacja))
Response: KONDYCJA
Model: C&RT

	Observed	Predicted stabilny	**Predicted zagrożony**	Row Total
Number	stabilny	413	4	417
Column Percentage		92.81%	12.50%	
Row Percentage		99.04%	0.96%	
Total Percentage		86.58%	0.84%	87.42%
Number	zagrożony	32	28	60
Column Percentage		7.19%	87.50%	
Row Percentage		53.33%	46.67%	
Total Percentage		6.71%	5.87%	12.58%
Count	All Groups	445	32	477
Total Percent		93.29%	6.71%	

Literatura

1. Altman E., Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy, The Journal of Finance, September 1968
2. Andreasik J., Salej A., Klasyfikacja zmiennych modeli prognozowania stanu zagrożenia upadłością przedsiębiorstw, Barometr Regionalny nr 1 (4) 2005, WSZiA w Zamościu
3. Cichosz P., Systemy uczące się, Wydawnictwa Naukowo-Techniczne, Warszawa 2000
4. Gruszczyński M., Modele i prognozy zmiennych jakościowych w finansach i bankowości, Monografie i Opracowania, SGH, Warszawa 2001
5. Hadasik D., Upadłość przedsiębiorstw w Polsce i metody jej prognozowania, AE w Poznaniu, Poznań 1998.
6. Koronacki J., Ćwik J., Statystyczne systemy uczące się Wydawnictwa Naukowo-Techniczne, Warszawa 2005
7. Krawiec K., Stefanowski J., Uczenie maszynowe i sieci neuronowe, Wydawnictwo Politechniki Poznańskiej, 2004
8. Lasek M., Data Mining. Zastosowania w analizach i ocenach klientów bankowych, Biblioteka Menedżera i Bankowca, Warszawa 2002.
9. materiały informacyjne z zastosowań systemu Statistica Data Miner ze strony www.statsoft.pl
10. Nowak E., Propozycje zmiennych oceniających zagrożenie dalszego funkcjonowania przedsiębiorstwa. Raport projektu „System przeciwdziałania powstawaniu bezrobocia na terenach słabo zurbanizowanych", Zamość 2006
11. praca zbiorowa Statystyka i Data Mining w praktyce – Statsoft Polska, Kraków 2004
12. Rutkowski A., Prognozowanie zagrożenia upadłością na podstawie sprawozdań finansowych, Nasz Rynek Kapitałowy nr 4/99.
13. Walanus A., Demski T., Data Mining – inteligencja biznesowa, MM Magazyn Przemysłowy 3/2004 s.38