PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
Nr 86 — 2009
Wizualizacja wyników badań marketingowych – podejścia, metody i zastosowania

**Mariusz Łapczyński**
Uniwersytet Ekonomiczny w Krakowie

# SPOSOBY WIZUALIZACJI DRZEW KLASYFIKACYJNYCH I REGRESYJNYCH CART

## 1. Drzewa klasyfikacyjne i regresyjne CART – wprowadzenie

Drzewa klasyfikacyjne CART [Breiman i in. 1984] to narzędzie analityczne *data mining*, które jest uznawane za najbardziej zaawansowaną metodę podziału rekurencyjnego. Mimo że metoda ta powstała na początku lat 80. ubiegłego wieku, to do dziś doczekała się tylko nieznacznych modyfikacji. Próbowano wprawdzie stworzyć bayesowski CART [Chapman, George, McCulloch 1998], dokonywano jego modyfikacji w NASA (pakiet IND) [Buntine 1992], usiłowano także udoskonalić podział drzew (FACT) przez połączenie właściwości CART i liniowej analizy dyskryminacyjnej [Loh, Vanichsetakul 1988], podejmowano także próby zastąpienia wielokrotnej walidacji krzyżowej metodą Monte Carlo [Crawford 1989], jednak rdzeń metody z jego nowatorskimi rozwiązaniami do dziś pozostał niezmieniony.

Celem artykułu jest prezentacja sposobów wizualizacji modeli drzew klasyfikacyjnych i regresyjnych zbudowanych za pomocą tego algorytmu. Poniższe zestawienie pokaże, w jaki sposób producenci oprogramowania „radzą" sobie z graficznym przedstawieniem struktur drzewa, które niejednokrotnie mają kilkadziesiąt liści i których wydruk na racjonalnym formacie papieru staje się niemożliwy.

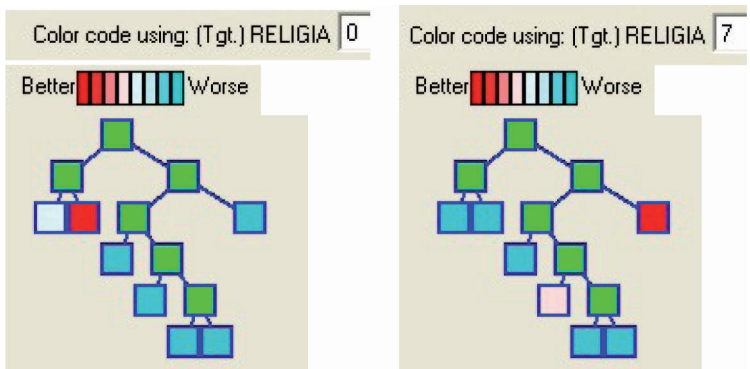
## 2. Wizualizacja modeli klasyfikacyjnych CART

Do budowy drzew klasyfikacyjnych wykorzystano plik „flagi" pobrany ze strony repozytorium *Machine Learning* Uniwersytetu w Kalifornii[1]. Zmienna zależna „religia" zawierała 8 wariantów, zaś zbiór 28 zmiennych niezależnych obejmował m.in.: kontynent (1 = Ameryka Północna, 2 = Ameryka Południowa, 3 = Europa,

[1] UCI Machine Learning Repository [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California, School of Information and Computer Science. W zbiorze obserwacji znajdują się 194 przypadki, z których każdy odnosi się do innego państwa, a dokładniej – do innej flagi narodowej. Zbiór zmiennych dotyczy charakterystyk tych krajów oraz charakterystyk flag.

4 = Afryka, 5 = Azja, 6 = Oceania), strefę wyznaczoną południkiem 0° i równikiem (1 = NE, 2 = SE, 3 = SW i 4 = NW), powierzchnię kraju [km$^2$], liczbę pionowych pasów na fladze, liczbę poziomych pasów na fladze, występowanie koloru żółtego lub złotego na fladze, liczbę symboli przedstawiających słońce lub gwiazdę, występowanie półksiężyca itp. Stosunkowo duża liczba wariantów zmiennej objaśnianej sprawia potencjalne problemy z czytelną wizualizacją modelu drzewa.

W programie CART® firmy Salford Systems (rys. 1) badacz może wybrać interesujący go wariant zmiennej zależnej (tu: „religia" = 0 i „religia" = 7) i automatycznie wszystkie liście drzewa zmienią kolor, w zależności od tego, jaki procent danej klasy zawierają. Odcienie czerwieni[2] oznaczają wysoki odsetek danej klasy, odcienie niebieskie zaś sytuację odwrotną.



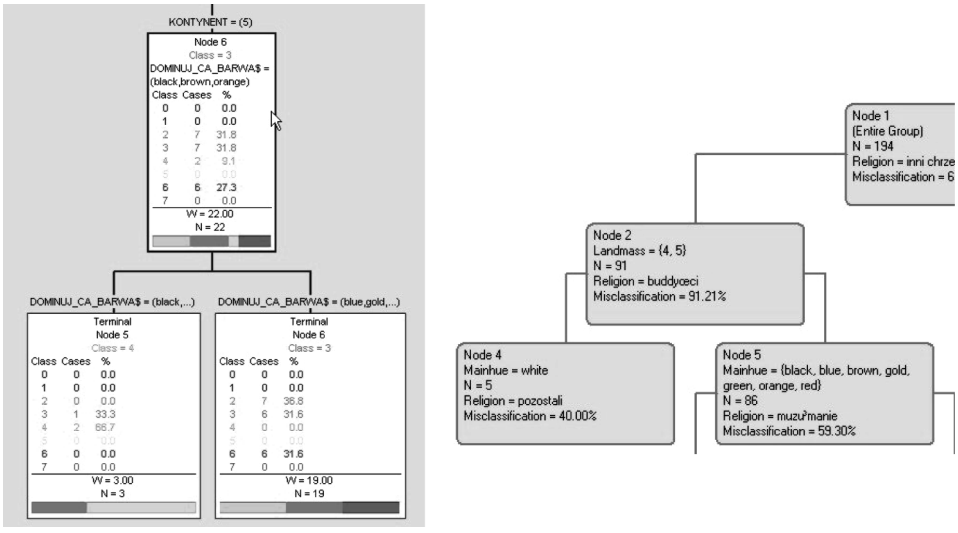

Rys. 1. Schemat drzewa klasyfikacyjnego z wyróżnionymi wariantami zmiennej zależnej

Źródło: opracowanie własne z wykorzystaniem programu CART pro EX V6.0.

Standardowy widok węzłów drzewa klasyfikacyjnego to: a) lista wariantów zmiennej objaśnianej wraz z informacją o częstości występowania każdego z nich (rys. 2 po lewej stronie); b) informacja o wariancie występującym najczęściej wraz z błędem klasyfikacji (rys. 2 po prawej stronie); c) wykres słupkowy zmiennej objaśnianej w węźle (rys. 3 po lewej stronie).

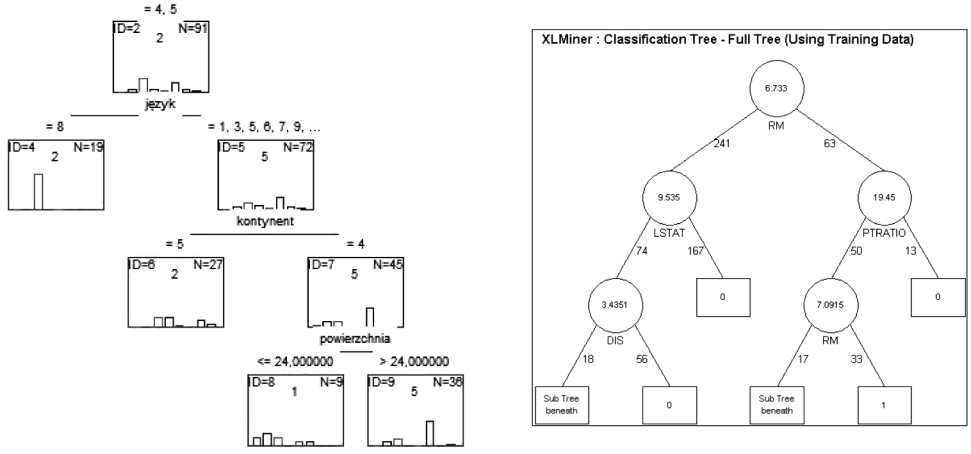
Jeśli chodzi o opis wariantów predyktora jakościowego lub wartości predyktora ilościowego, to zwykle informacja ta znajduje się nad węzłami potomnymi. Wyjątkiem jest tu program DTREG, w którym opis ten zamieszczono wewnątrz węzłów, co pozwala na pełną wizualizację. Ciekawy pomysł prezentacji całej struktury drzewa ma program XLMiner, gdzie wykres złożonego modelu jest podzielony na kilka mniejszych rysunków.

[2] Wydruk w odcieniach szarości nie pozwala niestety na pełną prezentację różnych sposobów wizualizacji drzew klasyfikacyjnych. Zainteresowanych czytelników odsyła się do wersji demonstracyjnych opisywanych tu programów komputerowych.




Rys. 2. Widok węzłów drzewa klasyfikacyjnego wykonanego w programie CART pro EX V6.0 (po lewej) i w programie DTREG (po prawej)

Źródło: opracowanie własne z wykorzystaniem programu CART pro EX V6.0 i programu DTREG.

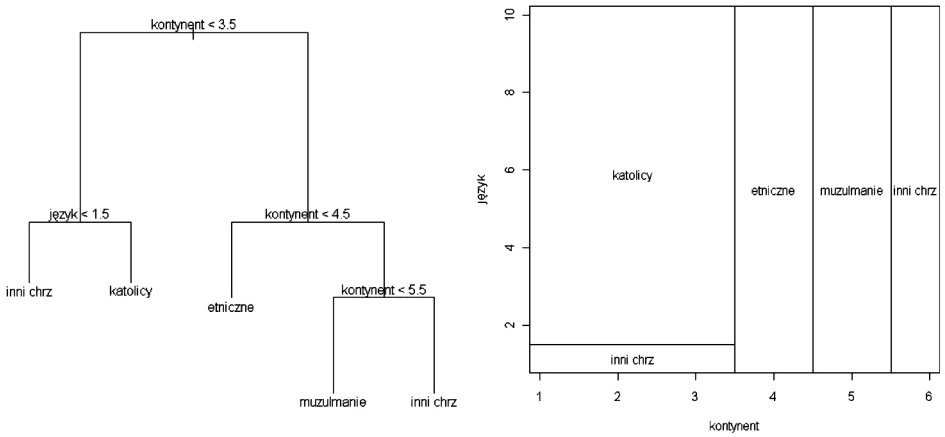



Rys. 3. Widok węzłów drzewa klasyfikacyjnego wykonanego w programie STATISTICA 8.0 (po lewej) i w programie XLMiner (po prawej)

Źródło: opracowanie własne z wykorzystaniem programu STATISTICA 8.0 i programu XLMiner.

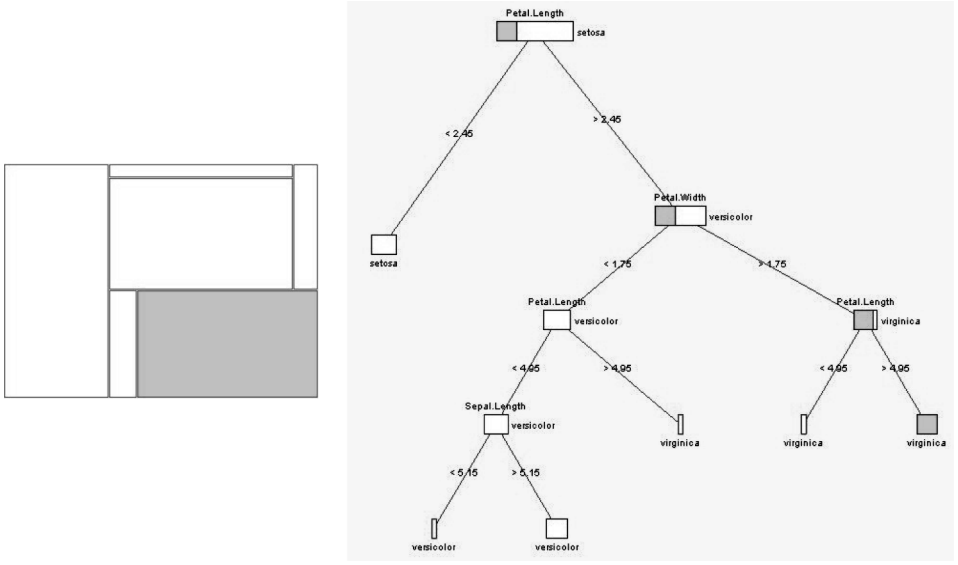
Nieco odmiennym sposobem prezentacji drzewa klasyfikacyjnego jest przedstawienie liści na wykresach przestrzennych. W programie R jest to możliwe wówczas, gdy w podziale drzewa uczestniczyły dwie zmienne niezależne (rys. 4), natomiast

w programie KLIMT można to robić dla dowolnego modelu (rys. 5). W tym drugim wypadku wielkość węzłów jest proporcjonalna do liczby zawartych w nich przypadków, natomiast wykresy przestrzenne nazywane są mapami drzewa (*tree maps*)[3].




Rys. 4. Struktura drzewa oraz podział przestrzeni dla liści drzewa w programie R

Źródło: opracowanie własne z wykorzystaniem programu R (pakiet *tree*).

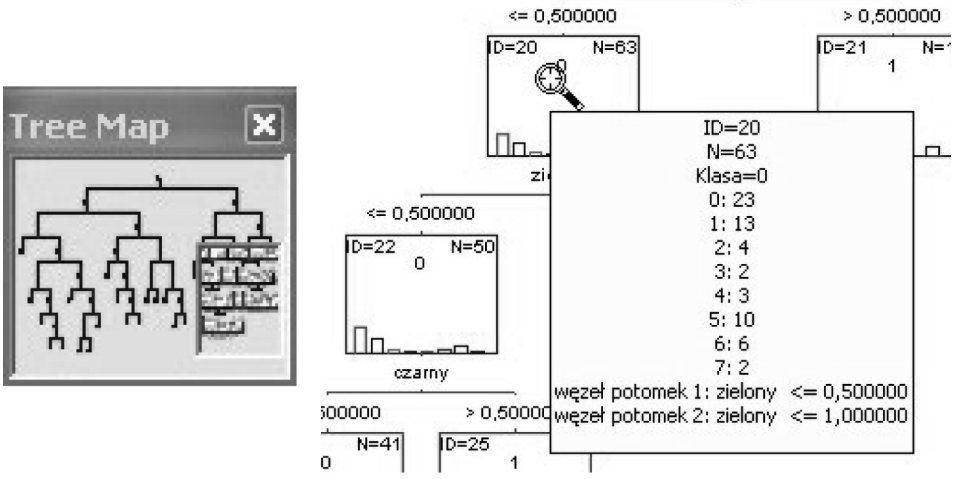



Rys. 5. Struktura drzewa (po prawej) i mapa drzewa (po lewej) w programie KLIMT

Źródło: opracowanie własne z wykorzystaniem programu KLIMT.

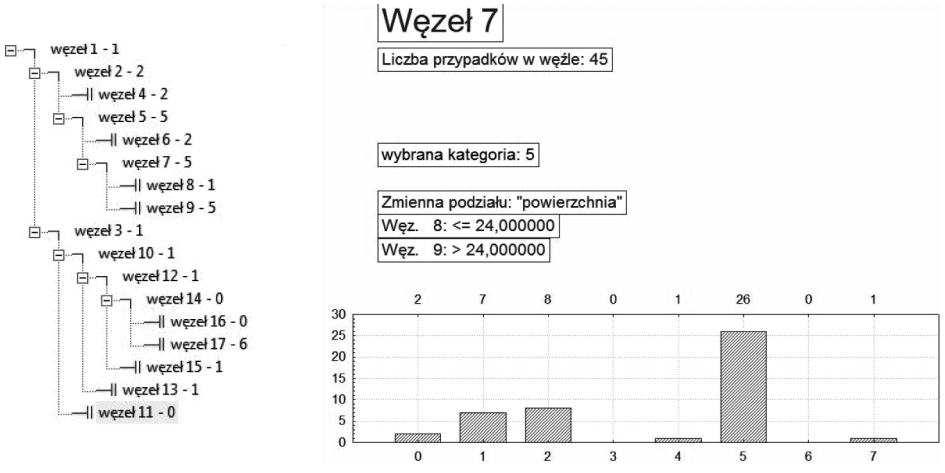
[3] Model klasyfikacyjny w programie KLIMT zbudowano wyjątkowo na podstawie danych z pliku „kwiaty irysa".

Przy bardzo złożonych strukturach drzewa twórcy oprogramowania proponują dwa podejścia. Pierwsze z nich pozwala na przegląd fragmentów modelu za pomocą tzw. mapy drzewa[4] lub za pomocą narzędzi graficznych typu „lupa" (rys. 6).




Rys. 6. Mapa drzewa umożliwiająca podgląd wybranego fragmentu modelu w programie CART pro EX V6.0 (po lewej) oraz narzędzie typu „lupa" w programie STATISTICA 8.0 (po prawej)

Źródło: opracowanie własne z wykorzystaniem programu CART pro EX V6.0 i programu STATISTICA 8.0.




Rys. 7. Alternatywny sposób prezentacji struktury drzewa w programie STATISTICA 8.0

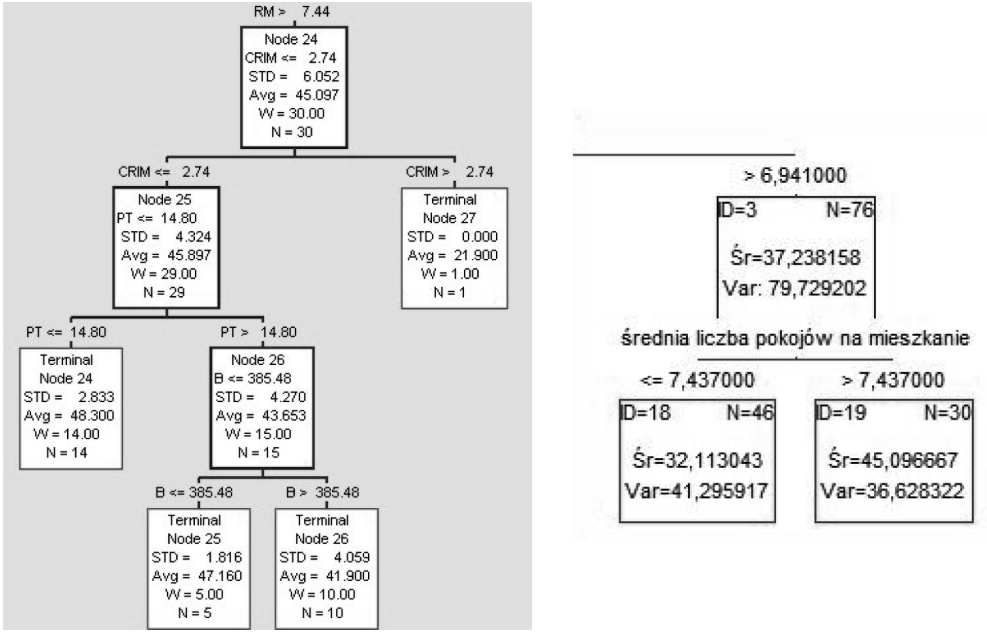
Źródło: opracowanie własne z wykorzystaniem programu STATISTICA 8.0.

[4] Termin *Tree Map* jest tutaj rozumiany inaczej niż w programie KLIMT.

Drugie z nich to szczegółowy podgląd poszczególnych liści drzewa (rys. 7). W pakiecie STATISTICA 8.0 wygląda to jak powszechnie znana struktura katalogów. Każdy z węzłów można dodatkowo przedstawić w postaci wykresu słupkowego, na którym znajdują się również informacje o ewentualnym dalszym podziale tego węzła.

## 3. Wizualizacja modeli regresyjnych CART

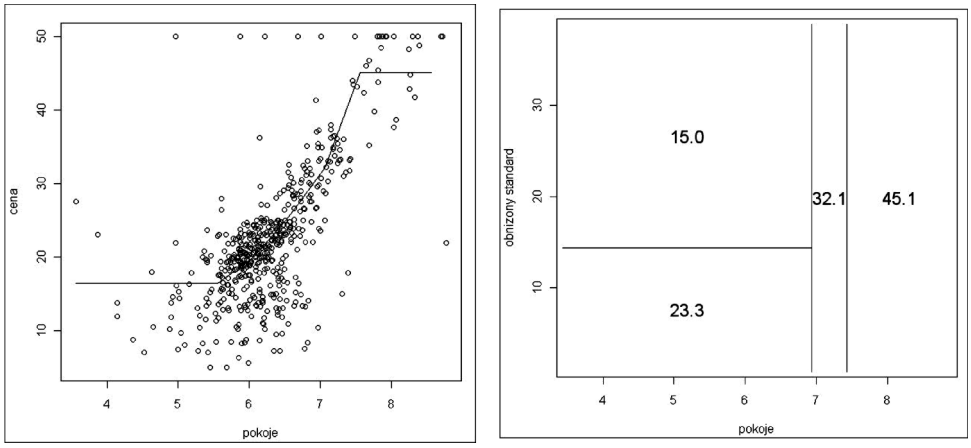
W przypadku drzew regresyjnych zmienna zależna jest ilościowa, co sprawia, że w jego liściach zamiast rozkładów procentowych znajduje się informacja o średniej i odchyleniu standardowym (ew. wariancji) bądź o medianie i odchyleniu przeciętnym (rys. 8). Wszystkie prezentowane tu modele powstały na podstawie popularnego zbioru obserwacji dotyczącego cen nieruchomości w Bostonie. Zmienna zależna to cena nieruchomości, zmienne niezależne zaś to m.in.: wskaźnik przestępczości, stężenie tlenku azotu w powietrzu, przeciętna liczba pokoi, dostęp do autostrady itp.




Rys. 8. Liście drzewa regresyjnego w programie CART pro EX V6.0 (po lewej) i drzewa regresyjnego w programie STATISTICA 8.0 (po prawej)

Źródło: opracowanie własne z wykorzystaniem programu CART pro EX V6.0 i programu STATISTICA 8.0.

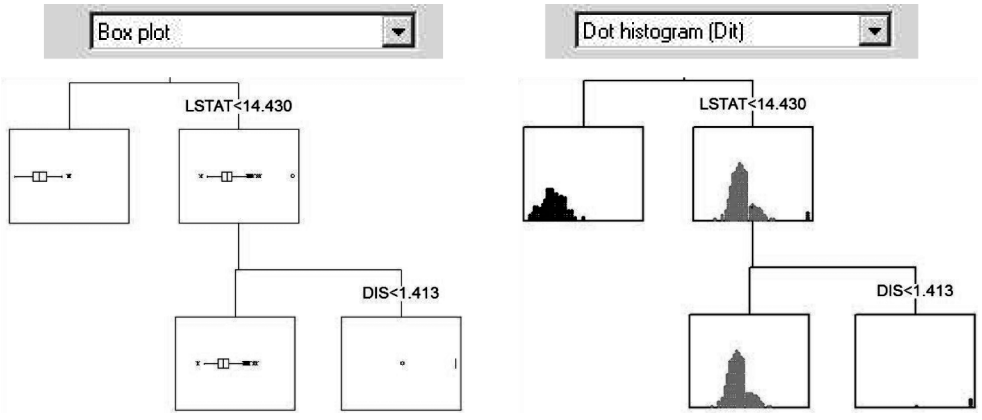
Program R, który niezbyt dobrze przedstawia złożone struktury drzewa z długimi opisami zmiennych niezależnych, pozwala na utworzenie funkcji schodkowych oraz interesujący podział przestrzeni zmiennych (rys. 9).




Rys. 9. Funkcja schodkowa (po lewej) i podział przestrzeni zmiennych (po prawej) dla drzewa regresyjnego w programie R

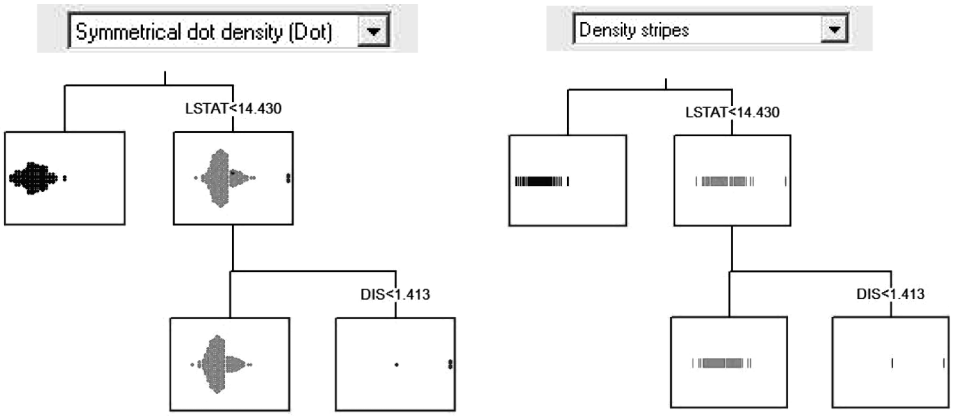
Źródło: opracowanie własne z wykorzystaniem programu R (pakiet *tree*).

W programie SYSTAT 11.0 wprowadzono z kolei stosunkowo nowatorski sposób wizualizacji liści. Wykorzystano fakt, że zmienna zależna jest ilościowa, co pozwoliło użyć wykresu ramka-wąsy, histogramu, „lustrzanego", histogramu i pasków gęstości (*density stripes*) – rys. 10 i 11.




Rys. 10. Wykres ramka-wąsy i histogram w liściach drzewa w programie SYSTAT

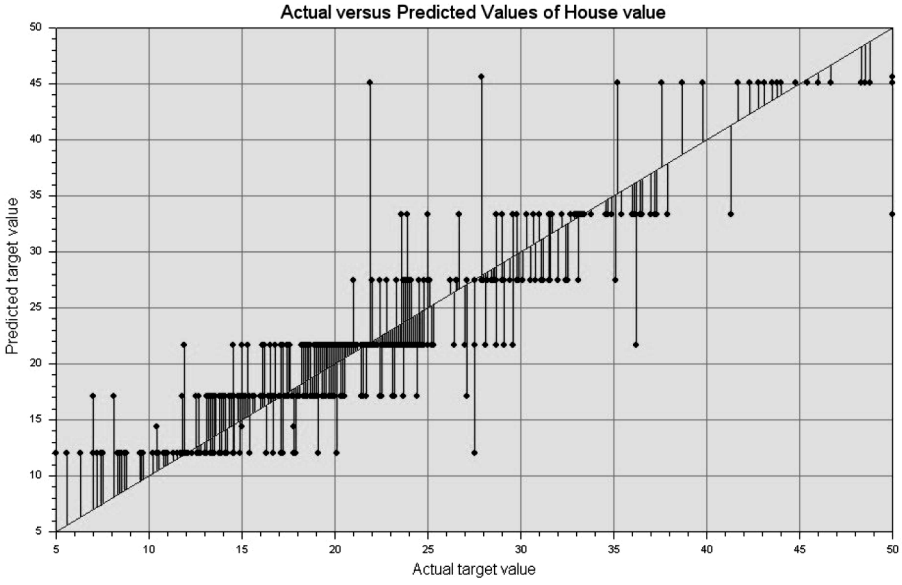
Źródło: opracowanie własne z wykorzystaniem programu SYSTAT 11.0.




Rys. 11. Lustrzane histogramy i paski gęstości w liściach drzewa w programie SYSTAT

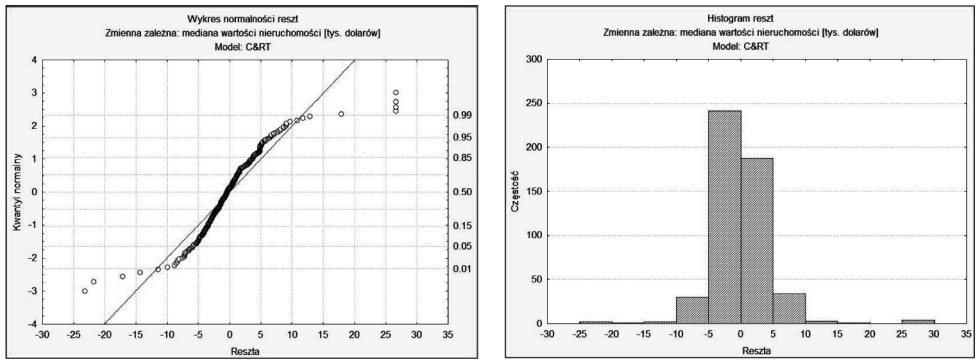
Źródło: opracowanie własne z wykorzystaniem programu SYSTAT 11.0.

Ocena „klasycznych" modeli regresyjnych obejmuje m.in. porównanie wartości obserwowanych z przewidywanymi oraz analizę reszt. Podobne narzędzia są także dostępne w programach, w których zaimplementowano algorytm CART, co ilustrują dwa kolejne rysunki (rys. 12 i 13).




Rys. 12. Wykres wartości obserwowanych i przewidywanych w programie DTREG

Źródło: opracowanie własne z wykorzystaniem programu DTREG.




Rys. 13. Analiza reszt w modelu regresyjnym

Źródło: opracowanie własne z wykorzystaniem programu STATISTICA 8.0.

## 4. Podsumowanie

Drzewa klasyfikacyjne i regresyjne są popularnym narzędziem *data mining* wykorzystywanym nie tylko na potrzeby analitycznego CRM, ale również w analizie danych ankietowych. Przewagą drzew nad innymi narzędziami analitycznymi, np. sieciami neuronowymi czy regresją logistyczną, jest czytelna graficzna prezentacja modelu. Zaleta ta zanika jednak w sytuacji, kiedy wynikiem analizy jest drzewo o dużej głębokości i wielkości, bowiem jego wydruk na papierze o rozsądnym formacie staje się niemożliwy. Z pomocą przychodzą tutaj rozwiązania zaproponowane przez producentów oprogramowania do analizy danych. Programy omówione w niniejszym artykule są jedynie kilkoma przykładami z szerokiej gamy dostępnych na rynku produktów, takich jak chociażby SPSS Clementine, IMB Intelligent Miner, Python 2.5 czy SAS Enterprise Miner. Spośród wymienionych powyżej na uwagę zasługują CART®, KLIMT i STATISTICA jako te dysponujące względnie nowatorskimi sposobami wizualizacji drzewa.

## Literatura

Breiman L., Friedman J., Olshen R., Stone C., *Classification and Regression Trees*, CRC Press, London 1984.

Buntine W., *Tree Classification Software*, „Technology 2002", Baltimore, December 1992.

Chapman H.A., George E.I., McCulloch R.E., *Bayesian CART Model Search*, „Journal of the American Statistical Association", September 1998 vol. 93, no 443, s. 935-960.

Crawford S.L., *Extension to the CART Algorithm*, „International Journal Man-Machine Studies" 1989 vol. 31, s. 197-217.

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

Loh W.-Y., Vanichsetakul N., *Tree-structured Classification Via Generalized Discriminant Analysis*, „Journal of the American Statistical Association", September 1988 vol. 83, no 403, s. 715-729.

Łapczyński M., *Drzewa klasyfikacyjne CART jako alternatywa dla klasycznych metod analizy danych marketingowych*, [w:] *Marketing*, D. Surówka-Marszałek (red.), Zeszyty Naukowe Krakowskiej Szkoły Wyższej im. Andrzeja Frycza Modrzewskiego, Kraków 2005, s. 135-145.

Łapczyński M., *Sposoby wizualizacji modeli drzew klasyfikacyjnych*, [w:] *Marketing*, D. Surówka-Marszałek (red.), Zeszyty Naukowe Krakowskiej Szkoły Wyższej im. Andrzeja Frycza Modrzewskiego, Kraków 2007, s. 189-198.

Walesiak M., Gatnar E. (red.), *Statystyczna analiza danych z wykorzystaniem programu R*, PWN, Warszawa 2009.

## WAYS OF VISUALIZING CART MODELS

### Summary

The goal of this article is to show different manners of visualizing classification and regression trees built with CART algorithm. In addition to final outcome – tree structure – the author demonstrates tree maps, diagrams of predicted and actual values, exploration of terminal nodes, partition of a tree, graphs of residuals and graphs of step function. The list of presented software includes CART®, STATISTICA, DTREG, KLIMT, XLMiner, SYSTAT and R (package *tree*).