Techniki eksploracji danych(projekt),

Dr hab. inż. Barbara Gładysz, Środa, 11^{15}-13^{00} TN, B-4, s. 433

Krzysztof Gużewski

Mikołaj Krymarys

Mateusz Guściora

Baza danych: Samochody(producent,modele, parametry, cena)

Projekt:

1. Badania Danych
2. Badanie zależności zmiennych. Wykres pudełko-wąsy.
3. Algorytm k-najbliższych sąsiadów
4. Eksploracja danych metodą k-średnich.
5. Drzewa klasyfikacyjne, decyzyjne, regresyjne

Badania danych

Pracowaliśmy na fragmencie bazy danych sprzedaży najczęściej sprzedawanych samochodów w roku 2012 na amerykańskim rynku samochodowym. Badanie dotyczyło głównie Ceny w tysiącach $, która jest naszą zmienną najważniejszą i jej zależności (o ile jest) i wpływach z innymi zmiennymi. Badaliśmy potencjalną zależność 4 metodami. Pierwszą metodą jest stworzenie wykresu pudełko-wąsy. Kolejną jest analiza algorytmu najbliższego sąsiada, dzięki której mogliśmy ustalić przewidywaną cenę (predykcja). Trzecią metodą była metoda eksploracji danych k-średnich. Zbudowaliśmy także drzewa decyzyjne/regresyjne a następnie porównaliśmy metody.

Wpływ na Cenę w tyś $ podejrzewaliśmy w niektórych zmiennych m.in. producent, konie mechaniczne, pojemność silnika oraz przy szerokości i długości pojazdu(podanych w calach). Zmienne znajdujące się w bazie danych są poniżej:

-Producent

-Model

-Sprzedaż w tyś

-Ponowna sprzedaż w tyś $

-Typ: Jest to zmienna jakościowa przyjmująca 2 wartości. Passenger-auto osobowe, Car- Suv-y.

-Cena w tysiącach $

-Pojemność silnika w cm3

-Konie mechaniczne

-Rozstaw osi: Jest to odległość między przednią i tylną osią samochodu podana w calach

-Szerokość_cale: Szerokość samochodu w calach

-Długość_cale

-Masa_pojazdu_tys_funty

-Pojemność_zbiornika_paliwa: w galonach

-Wydajność_paliwowa: miles per galon

-Ostatnia_data_wydania

-Wspolczynnik_mocy_silnika

-Przewidywana_wartość_KNN

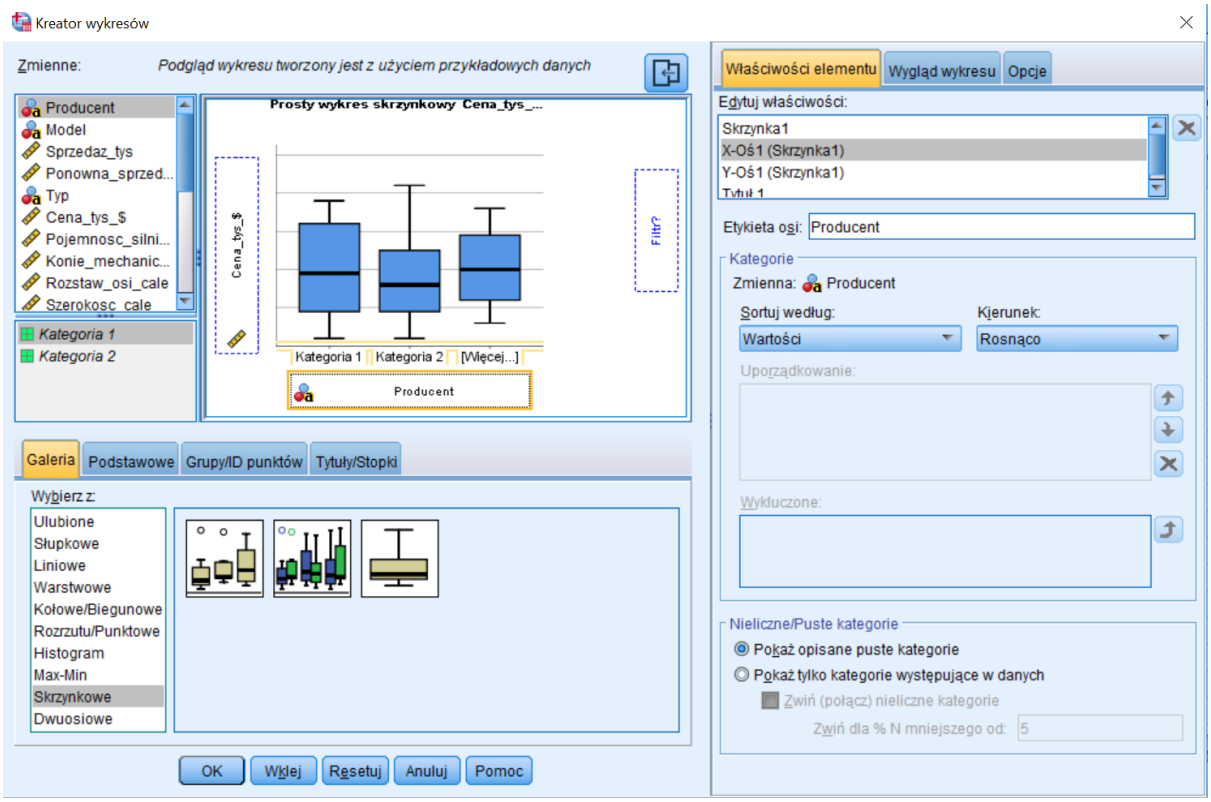
-KNN: przewidywana zmienna metodą najbiższych sąsiadów

-Blad_Mape

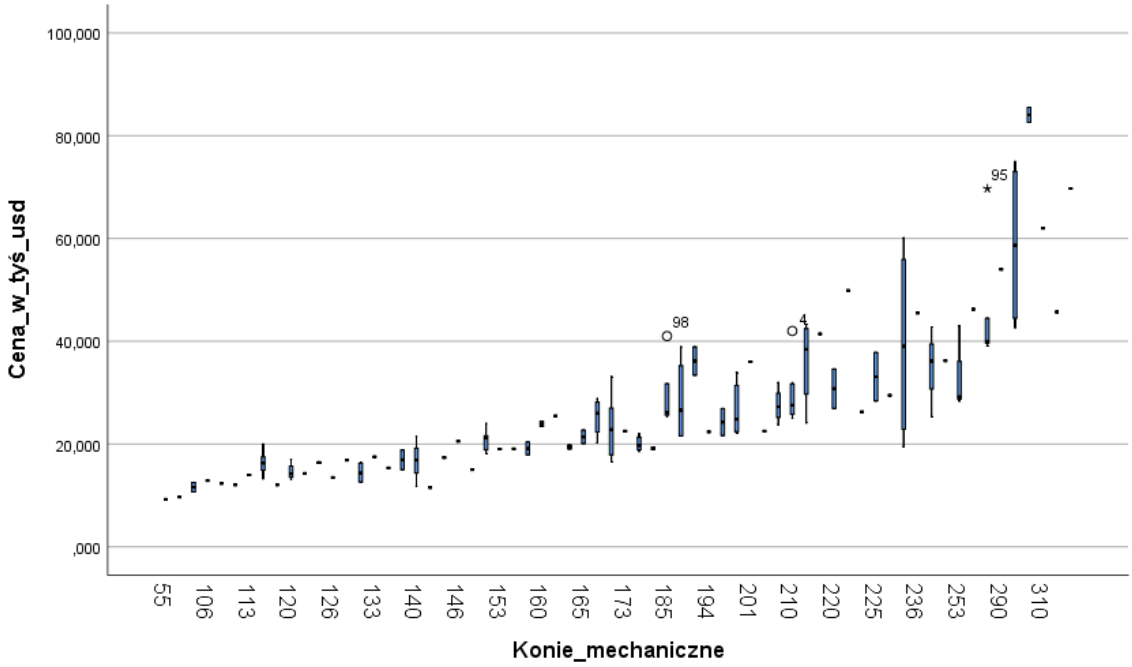
-Sredni_Blad_Mapy

1. Badanie zależności zmiennych. Wykres pudełko-wąsy.
   1.1. Opis metody
   1.2. Sposób użycia metody w SSPS
   1.3. Wyniki i ich interpretacja

Aby stworzyć wykres w SPSS należy wejść w zakładkę wykres→ kreator wykresu a następnie wybrać interesujący nas rodzaj wykresu oraz zmienne.



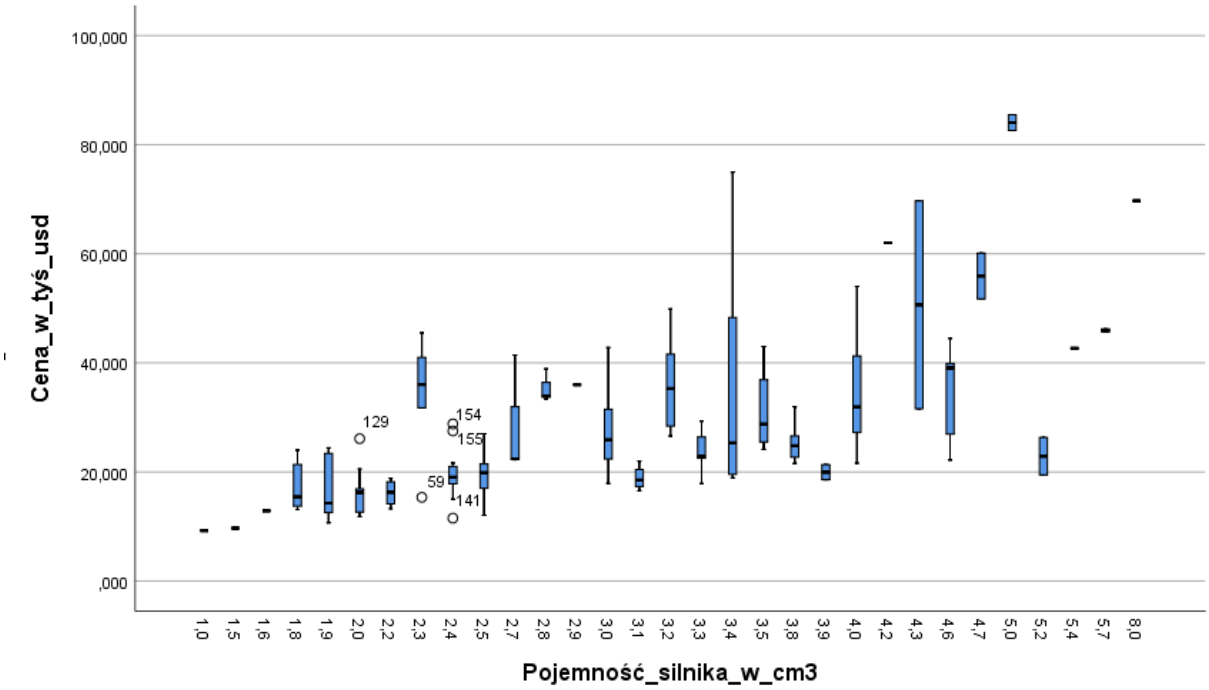
- Wykres zależności ceny od liczby koni mechanicznych.



Na wykresie możemy zauważyć 2 obserwacje nie typowe – dla zmiennej numer 4 i 98. Wraz ze wzrostem mocy, rośnie również cena. Poniżej 110 koni mechanicznych i powyżej 310 jest niewiele obserwacji z powodu małej popularności takich aut.

Obserwacją odstającą (przy ok. 70 000$ i ok. 290 koni mechanicznych) jest mercedes-B s klasy. Jego cena jest podobna do aut powyżej 300 koni mechaniczny. Jego cena jest wyższa niż innych samochodów w tym przedziale koni mechanicznych lecz ze względu na marka oraz inne parametry jest on wyżej ceniony.

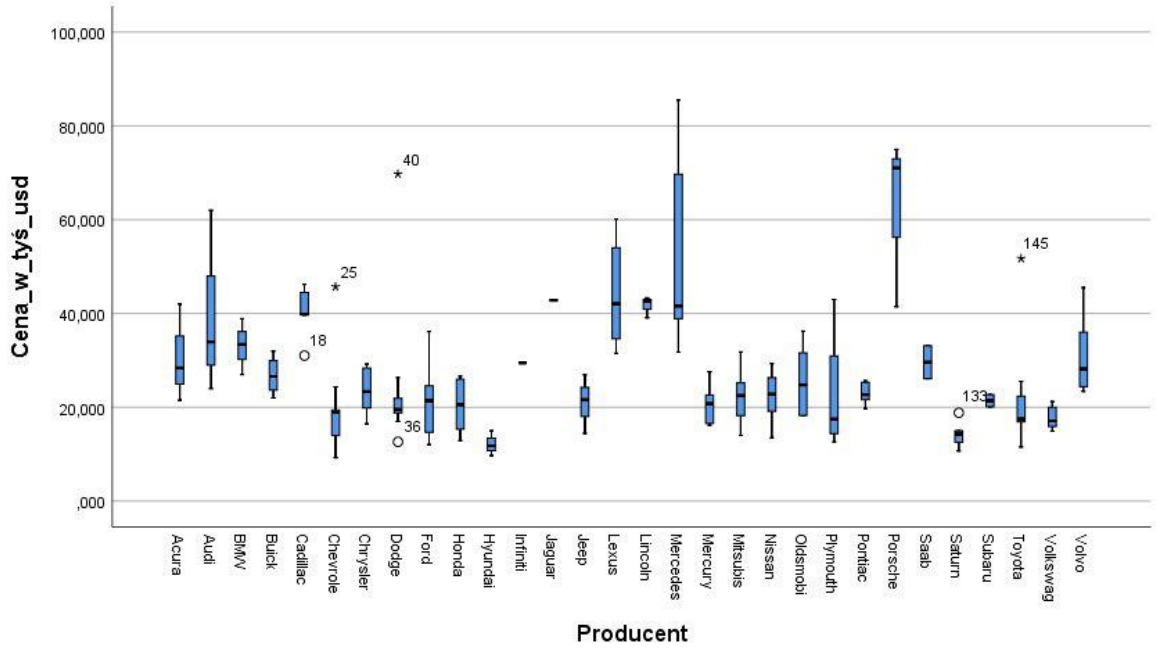
- Wykres zależności ceny od pojemności silnika w cm3.



Na wykresie możemy zauważyć 5 obserwacji podejrzanych o nietypowość – dla zmiennych o numerach: 59, 129, 141, 154, 155.
Wraz ze wzrostem pojemności silnika, rośnie również cena.

Brak obserwacji odstających. Istnieją jedynie obserwacje podejrzane o to.

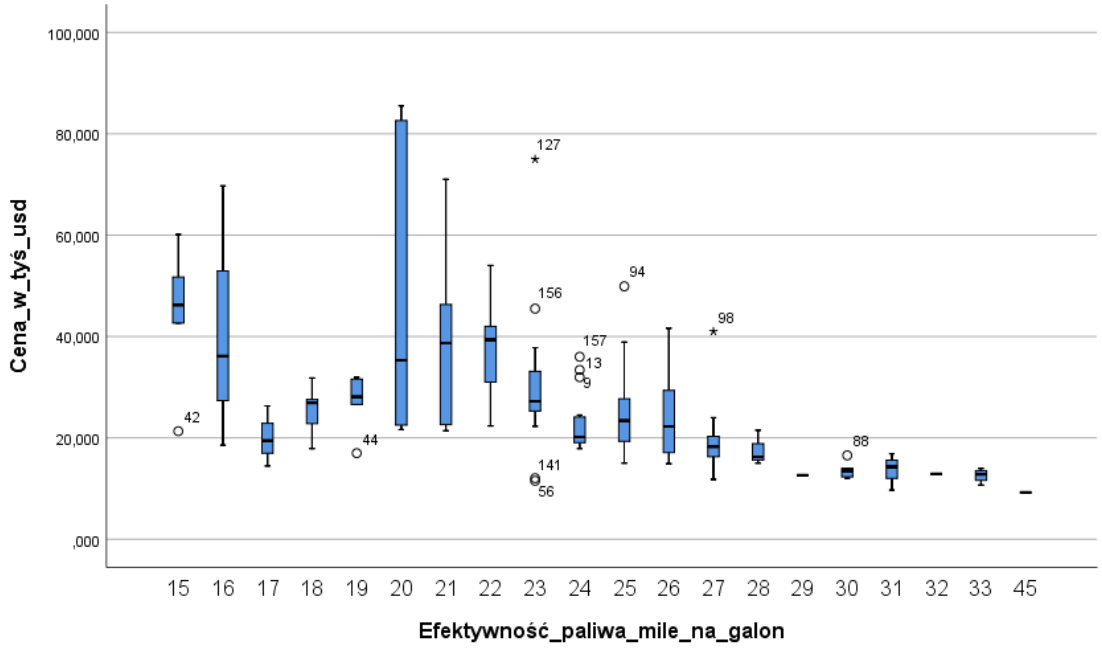
- Wykres zależności Ceny od producenta samochodu.



Na wykresie możemy zauważyć 3 obserwacje podejrzane o nietypowość – dla zmiennych o numerach: 18, 36, 133.
Najwyższą medianę ceny posiada producent Porsche, jednak to samochód od Mercedesa ma najwyższą cenę.
Najniższą medianę ceny posiada producent Hyundai, zaś najtańszy samochód należy do marki Chevrolet.
Na wykresie możemy również dostrzec producentów samochodów typu premium, których ceny są wyraźnie wyższe.

Obserwacje odstające dla marki Chevrolet. Większość samochodów produkowanych przez tego producenta mieści się w przedziale 10-25 tyś $ lecz istnieją pewne obserwacje odstające (droższe). Wynika to z kilku modeli bardziej luksusowych producenta.
Dla marki Dodge i Toyota jest podobnie. Dodge produkuje bardziej budżetowe(ok. 20 000) samochody ale zdarzają się też te o wyższej cenie (70 000$). Toyota w większości także produkuje tańsze (ok. 20 000) ale i niewiele droższych (50 000$).

- Wykres zależności ceny od efektywności paliwa(w milach/galony) tzn. długości na pojemność.



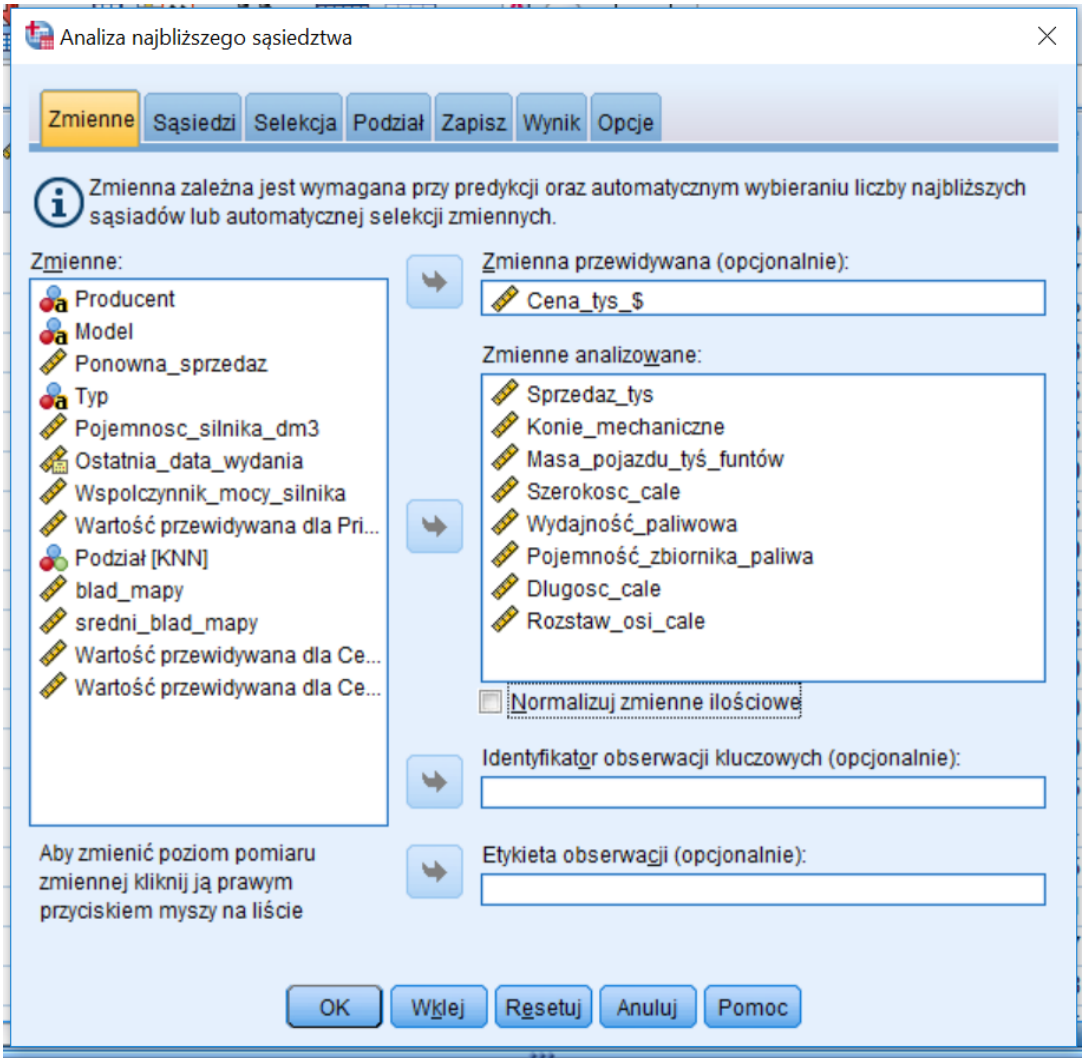
Wyższa cena wcale nie oznacza większej efektywności. Najwyższą efektywność mają samochody oscylujące wokół ceny 20 000$.
Obserwacje odstające pojawiają się przy efektywności paliwa dla 27 milach/galon. Oznacza to, że dla efektywności 27 mil/galon większość samochodów plasuje się w cenie ok. 20 000 $ ale jest też egzemplarz samochodu (mercedes, model SLK230), który dla tej efektywności przekracza cene 40 000$.

Algorytm k-najbliższych sąsiadów

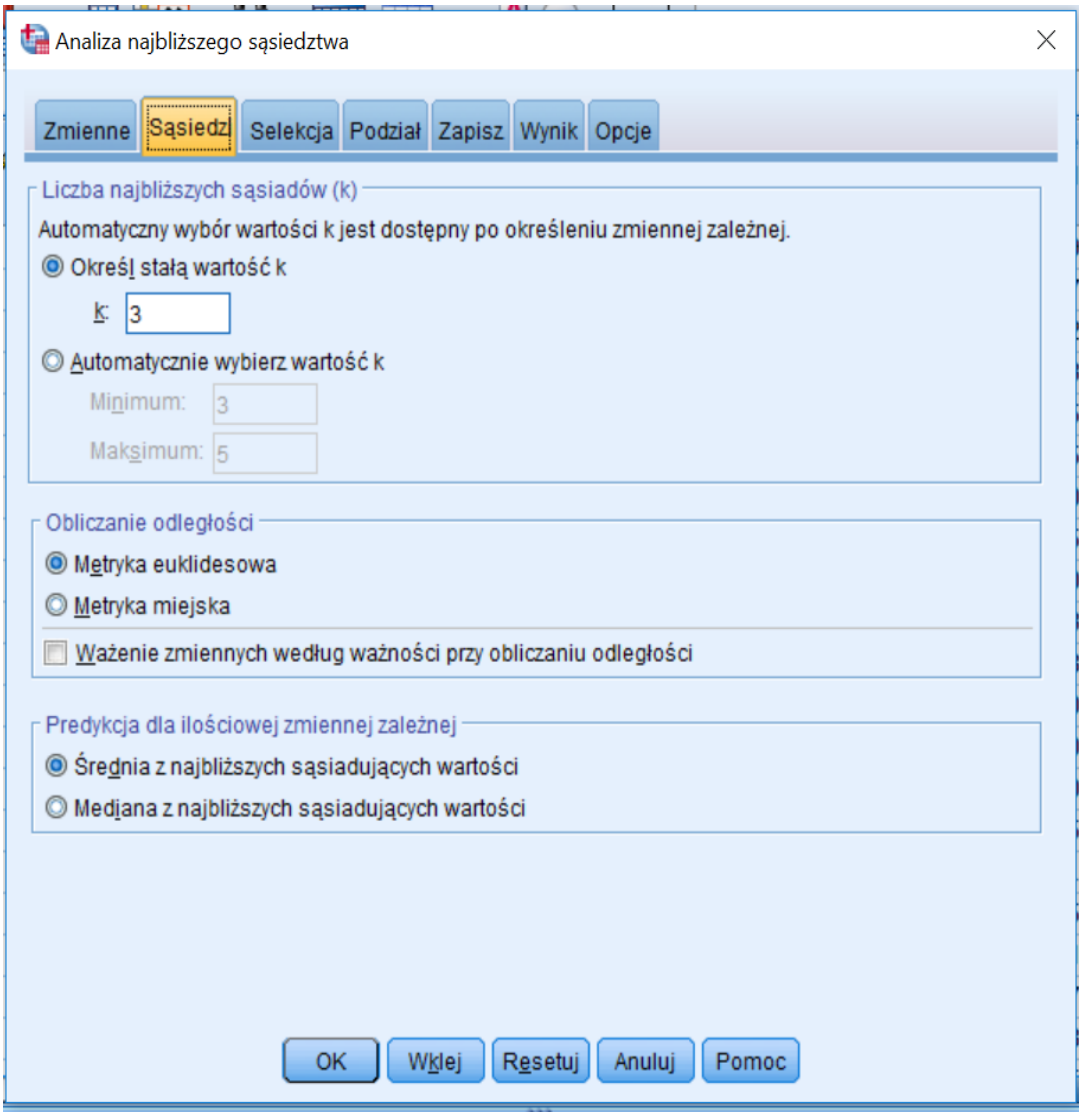
1. Definicja

Algorytm k-najbliższych sąsiadów służy do klasyfikacji zmiennych ale można zastosować go też do szacowania oraz przewidywanie zmiennych, opiera się na idei przewidywania nieznanych wartości poprzez dopasowanie ich do najbardziej podobnych znanych wartości. (k-ilu sąsiadów bierzemy pod uwagę może być dowolne)

2. Użycie metody w SPSS:



Żeby dotrzeć do wyżej przedstawionego okna, musimy na pasku zadać wejść w Analiza → Klasyfikacja → Najbliższego sąsiedztwa. W karcie Zmienne definiujemy zmienną przewidywaną (cena w tys. $), a w zmiennych analizowanych podajemy zmienne których wartości pozwolą nam na zbadanie najbliższych sąsiadów, a następnie przewidzeniu wartości głównej zmiennej. Możemy je normalizować (obróbić dane w celu umożliwienia ich wzajemnego porównywania i dalszej analizy).



W karcie Sąsiedzi zaznaczamy ilu sąsiadów ma znaleźć program oraz czy odległość ma liczyć metryką euklidesową czy miejską. W karcie Zapisz wybieramy dane które chcemy, żeby zostały zapisane jako kolejna zmienna.

3. Wynik i interpretacja

Poniższa tabela przedstawia fragment (15 wierszy) przewidywanej zmiennej celu, dla porównania cenę podaną w bazie danych oraz policzony błąd MAPE (Średni bezwzględny błąd procentowy jaki popełniamy przy szacowaniu wartości)

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{y_t - y_t^P}{y_t}\right| * 100\% \qquad t = 1,2,3,\ldots,n$$

Nr wiersza	cena($)	przewidywana cena($)	błąd MAPE	błąd średni
1	21500	15836	0,26344	0,1232239
2	28400	25058	0,11768	
3	-	-	-	
4	42000	39443	0,06088	
5	23990	21769	0,09258	
6	33950	28817	0,15119	
7	62000	55895	0,09847	
8	26990	27566	0,02134	
9	33400	32464	0,02802	
10	38900	35417	0,08954	
11	21975	22293	0,01447	
12	25300	30637	0,21095	
13	31965	27148	0,15070	
14	27885	27440	0,01596	
15	39895	50788	0,27304	

Poniższa tabela przedstawia nam 14 różnych kombinacji z parametrami K (od 2 do 5), normalizacją (Tak/Nie), metryką (euklidesowa/miejsca), ważone (Tak/Nie (mówi nam o tym czy zmienne były ważone według ważności przy obliczaniu odległości)

K	norm	metryka	ważone?	zmienne istotne	Wartość funkcji kryterium
2	Tak	Euklidesowa	Nie	sprzedaż w tys. konie mechaniczne masa własna	4468.3
2	Nie	Euklidesowa	Nie	sprzedaż w tys. konie mechaniczne masa własna szerokość	5 928,0
2	Tak	Euklidesowa	Tak	sprzedaż w tys. konie mechaniczne masa własna szerokość wydajnosc paliwa	3748.1
2	Nie	Euklidesowa	Tak	sprzedaż w tys. konie mechaniczne masa własna pojemnosc silnika	7323.5
4	Nie	Euklidesowa	Tak	sprzedaż w tys. konie mechaniczne pojemnosc silnika rozstaw osi pojemnosc zbiornika	5277.4
3	Tak	Euklidesowa	Tak	sprzedaż w tys. konie mechaniczne szerokosc	3481.8
5	Tak	Euklidesowa	Nie	sprzedaż w tys. konie mechaniczne masa	4160.3
5	Nie	Euklidesowa	Nie	sprzedaż w tys. konie mechaniczne rozstaw osi pojemnosc zbiornika szerokosc	6489.9
4	Nie	Miejska	Nie	konie mechaniczne wydajnosc pojemnosc zbiornika pojemnosc silnika	5 351,0
2	Tak	Miejska	Nie	konie mechaniczne masa	5025.8
3	Tak	Miejska	Nie	konie mechaniczne masa sprzedaz w tys.	4636.4
2	Tak	Miejska	Tak	konie mechaniczne dlugosc sprzedaz w tys. rozstaw osi	4057.2
4	Nie	Miejska	Tak	konie mechaniczne pojemnosc zbiornika sprzedaz w tys. rozstaw osi	6008.2
5	Nie	Miejska	Tak	konie mechaniczne pojemnosc silnika sprzedaz w tys. rozstaw osi wydajnosc masa	5706.4

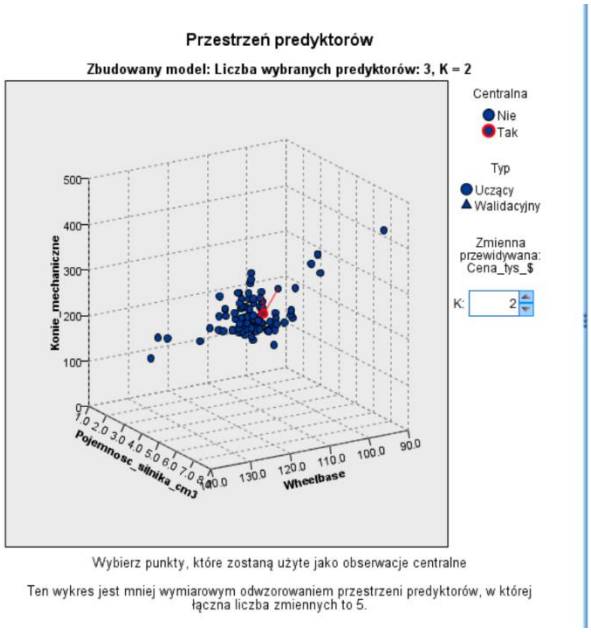
Algorytm k-najbliższych sąsiadów w spps, przykład (normalizacja-TAK, metryka euklidesowa, dla k=2)

Przewidywanie zmiennej celu: Cena_w_tyś_$

Analiza najbliższego sąsiedztwa (KNN)

Informacja o analizowanych danych

		N	Procent
Próba	Uczący	107	68.6%
	Wariant kontrolny	49	31.4%
Ważnych		156	100.0%
Wykluczone		1	
Łącznie		157	



K najbliższych sąsiadów i odległości

Wyświetlane dla wstępnych obserwacji centralnych

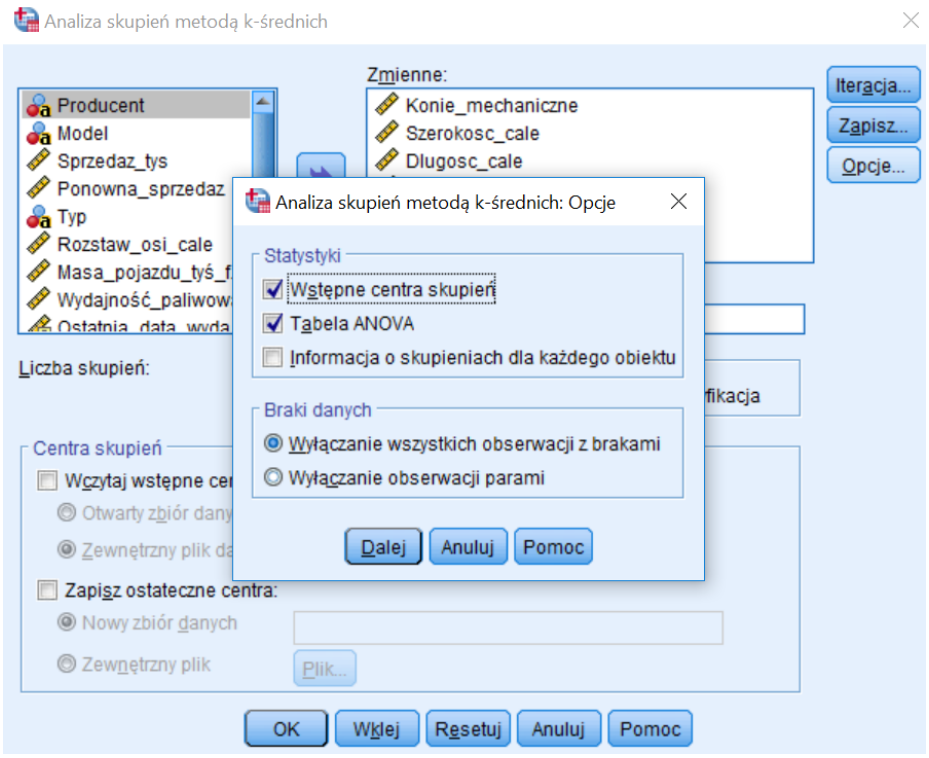
Obserwacja centralna	Najbliższe sąsiedztwo		Najmniejsze odległości	
	1	2	1	2
145	16	17	0.361	0.379

Wyciągając wnioski, średni błąd MAPE (12%) jest w granicach ufności. Najczęściej wybieranymi predyktorami były konie mechaniczne. Z tego względu możemy wyciągnąć wniosek, że wraz ze wzrostem koni mechanicznych w samochodzie, rośnie jego cena. Powinniśmy stosować metrykę miejską, ponieważ nie uwzględnia ona zmiennej „sprzedaż w tyś.” jako ważną. I słusznie, wiemy że cena wcale nie jest silnie uzależniona od wielkości sprzedaży.

Eksploracja danych metodą k-średnich.

Metoda k-średnich jest metodą należacą do grupy algorytmów analizy skupień tj. analizy polegającej na szukaniu i wyodrębnianiu grup obiektów podobnych (skupień). Algorytm ten tworzy centra, które skupiają pewien podzbiór(najbliższych rekordów).

Metoda k-średnich w programie spss. Wybieramy w menu Analiza→Klasyfikacja→Analiza skupień metodą k-średnich, w opcjach→ Wstępne,Tabela Anova.



Dla K=2

Analiza wariancji

	Skupienie		Błąd			
	Średni kwadrat	df	Średni kwadrat	df	F	Istotność
Sprzedaż_w_tyś	458559.871	1	1710.443	154	268.094	.000
Pojemność_silnika_w_cm3	.195	1	1.097	154	.178	.674
Konie_mechaniczne	16393.868	1	3129.349	154	5.239	.023
Szerokość_w_calach	3.511	1	11.970	154	.293	.589
Długość_w_calach	319.218	1	179.511	154	1.778	.184

Poziom istotności jest wysoki przy zmiennych Pojemność_silnika_w_cm3, Szerokość_w_calach, Długość_w_calach.

Dla K=3

Analiza wariancji

	Skupienie		Błąd			
	Średni kwadrat	df	Średni kwadrat	df	F	Istotność
Sprzedaż_w_tyś	238290.938	2	1603.832	153	148.576	.000
Pojemność_silnika_w_cm3	21.563	2	.824	153	26.178	.000
Konie_mechaniczne	134595.147	2	1497.538	153	89.878	.000
Szerokość_w_calach	76.736	2	11.068	153	6.933	.001
Długość_w_calach	685.010	2	173.816	153	3.941	.021

Poziom istotności jest bardzo mały co oznacza, że możemy brać wyniki jako istotne statystycznie.

Dla K=4

Analiza wariancji

	Skupienie		Błąd			
	Średni kwadrat	df	Średni kwadrat	df	F	Istotność
Sprzedaż_w_tyś	197515.458	3	851.459	152	231.973	.000
Pojemność_silnika_w_cm3	15.894	3	.799	152	19.888	.000
Konie_mechaniczne	91460.968	3	1473.228	152	62.082	.000
Szerokość_w_calach	86.371	3	10.446	152	8.268	.000
Długość_w_calach	980.842	3	164.614	152	5.958	.001

Poziom istotności jest bardzo mały co oznacza, że możemy brać wyniki jako istotne statystycznie.

Dla K=5

Analiza wariancji

	Skupienie		Błąd			
	Średni kwadrat	df	Średni kwadrat	df	F	Istotność
Sprzedaż_w_tyś	146023.412	4	913.076	151	159.925	.000
Pojemność_silnika_w_cm3	22.946	4	.512	151	44.783	.000
Konie_mechaniczne	96452.526	4	745.056	151	129.457	.000
Szerokość_w_calach	146.898	4	8.340	151	17.614	.000
Długość_w_calach	1618.586	4	142.315	151	11.373	.000

Poziom istotności jest bardzo mały co oznacza, że możemy brać wyniki jako istotne statystycznie.

Dla K=6

Analiza wariancji

	Skupienie		Błąd			
	Średni kwadrat	df	Średni kwadrat	df	F	Istotność
Sprzedaż_w_tyś	130284.549	5	470.302	150	277.023	.000
Pojemność_silnika_w_cm3	17.441	5	.546	150	31.927	.000
Konie_mechaniczne	74221.289	5	848.048	150	87.520	.000
Szerokość_w_calach	97.161	5	9.074	150	10.708	.000
Długość_w_calach	1428.715	5	138.802	150	10.293	.000

Poziom istotności jest bardzo mały co oznacza, że możemy brać wyniki jako istotne statystycznie.

Dla K=7

Analiza wariancji

	Skupienie		Błąd			
	Średni kwadrat	df	Średni kwadrat	df	F	Istotność
Sprzedaż_w_tyś	108210.779	6	487.943	149	221.769	.000
Pojemność_silnika_w_cm3	16.661	6	.464	149	35.883	.000
Konie_mechaniczne	67334.537	6	632.929	149	106.386	.000
Szerokość_w_calach	89.015	6	8.811	149	10.103	.000
Długość_w_calach	1502.097	6	127.190	149	11.810	.000

Poziom istotności jest bardzo mały co oznacza, że możemy brać wyniki jako istotne statystycznie.

Dla K=8

Analiza wariancji

	Skupienie		Błąd			
	Średni kwadrat	df	Średni kwadrat	df	F	Istotność
Sprzedaż_w_tyś	94729.641	7	397.707	148	238.190	.000
Pojemność_silnika_w_cm3	14.439	7	.460	148	31.389	.000
Konie_mechaniczne	58921.964	7	580.134	148	101.566	.000
Szerokość_w_calach	82.655	7	8.570	148	9.645	.000
Długość_w_calach	1236.576	7	130.458	148	9.479	.000

Na podstawie wyników otrzymanych w SPSS stworzyliśmy tabelkę zawierającą: liczbę klas, zmienne istotne, FC(funkcje celu). Funkcja celu liczona jest ze wzoru BCV/WCV. Według kryterium homogeniczności BCV/WCV → max tzn. badamy funkcje celów dla różnych k dopóki wartość tej funkcji nie maleje. FC zostało obliczone ze wzoru: $\frac{D}{d1^2+d2^2+\cdots+dn\hat{}2}$ (D-średnia odległości, suma

kwadratów odległości). FC→ max oznacza to, że badamy dopóki wartość FC rośnie gdy przestaje oznacza to że otrzymaliśmy największy wynik. Poziom istotności(alfa) żeby obserwacja była istotny statystycznie powinien wynosić <0,1.

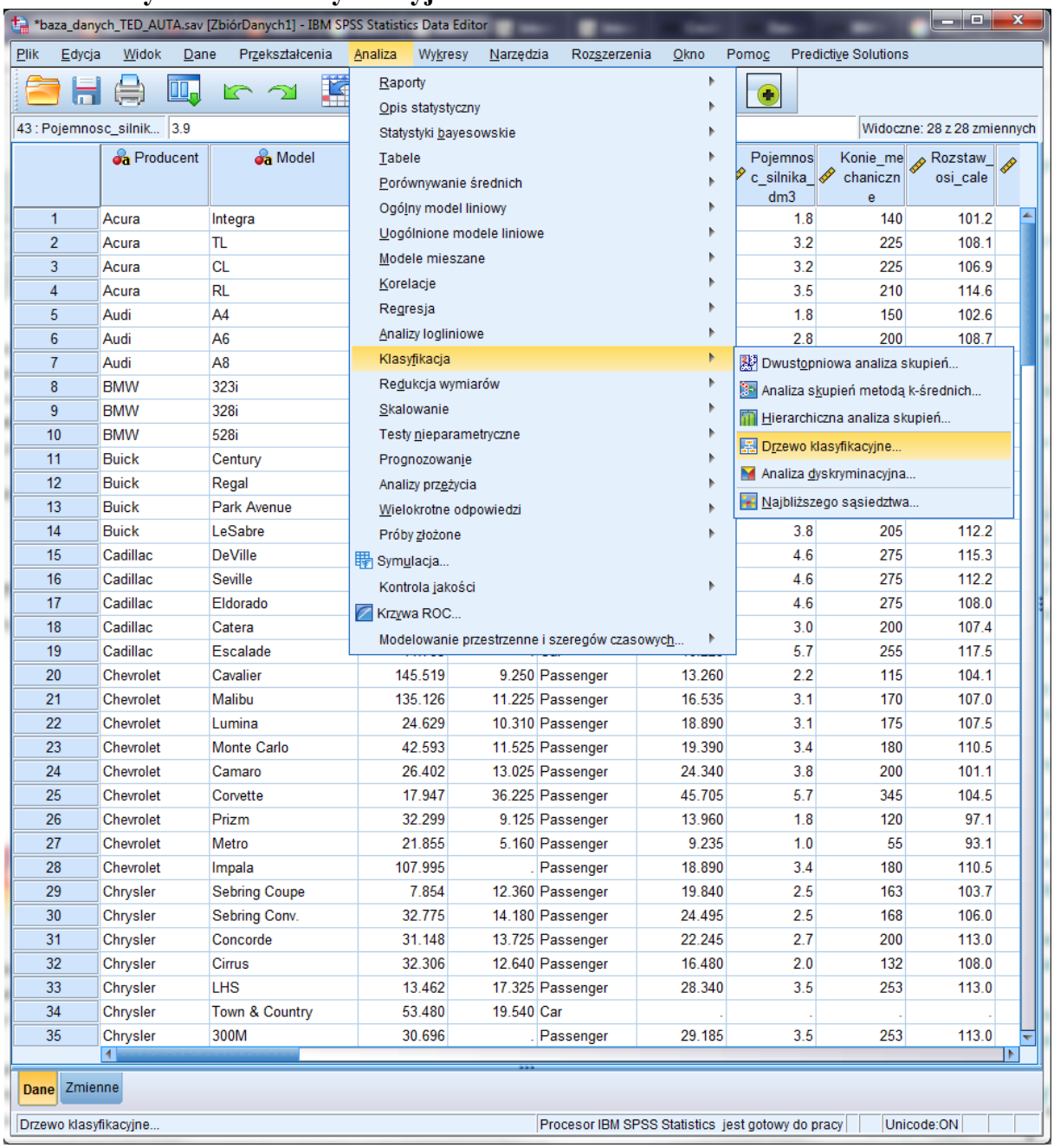
Przy bazie danych autu i ich parametrów przy k równym 7 funkcja celów jest największa.

K	Istotne zmienne	FC
2	Sprzedaż_w _tyś Konie_mechaniczne	0,02008
3	Sprzedaż_w _tyś Pojemność_silnika_w_cm3 Konie_mechaniczne Szerkość_w_calach Długość_w_calach	0,023453
4	Sprzedaż_w _tyś Pojemność_silnika_w_cm3 Konie_mechaniczne Szerkość_w_calach Długość_w_calach	0,045323
5	Sprzedaż_w _tyś Pojemność_silnika_w_cm3 Konie_mechaniczne Szerkość_w_calach Długość_w_calach	0,048282
6	Sprzedaż_w _tyś Pojemność_silnika_w_cm3 Konie_mechaniczne Szerkość_w_calach Długość_w_calach	0,049616
7	Sprzedaż_w _tyś Pojemność_silnika_w_cm3 Konie_mechaniczne Szerkość_w_calach Długość_w_calach	0,055765
8	Sprzedaż_w _tyś Pojemność_silnika_w_cm3 Konie_mechaniczne Szerkość_w_calach Długość_w_calach	0,054613

Drzewa klasyfikacyjne, decyzyjne, regresyjne

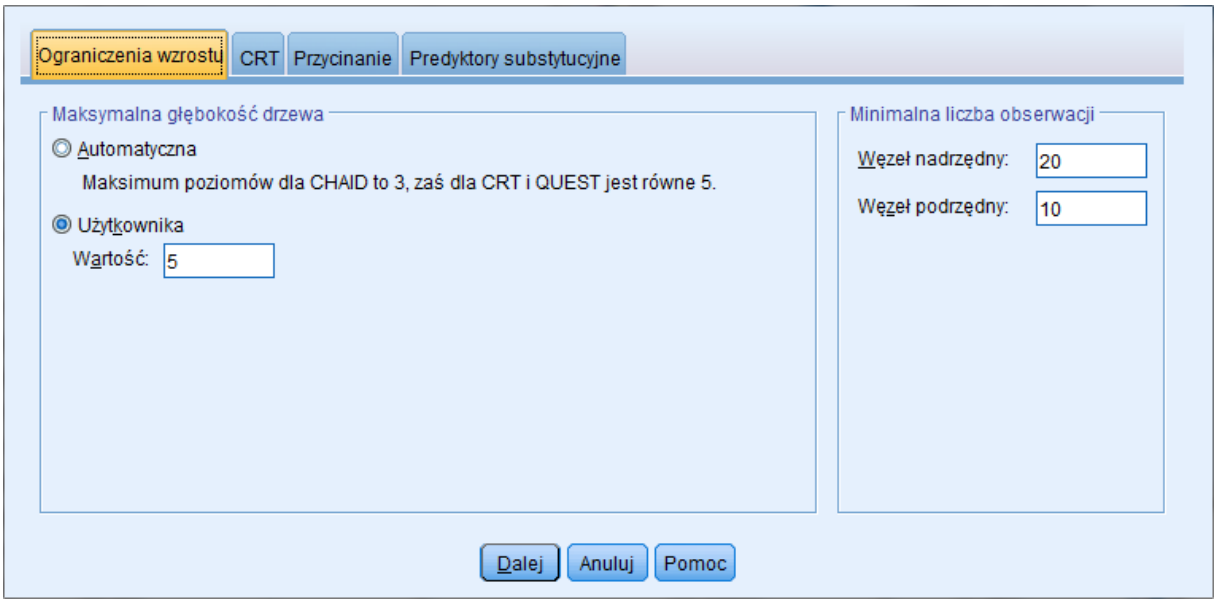
Drzewo decyzyjne jest zbiorem węzłów decyzyjnych połączonych za pomocą gałęzi, rozchodzących się w dół od korzenia aż do kończących liści. Zaczynając od korzenia, który zwyczajowo umieszczany jest na górze schematu decyzyjnego, atrybuty są sprawdzane w węzłach decyzyjnych, a każde możliwe wyjście zaznaczane jako gałąź. Każdą gałąź prowadzi albo do innego węzła decyzyjnego, albo do liścia.

Jak stworzyć drzewo klasyfikacyjne

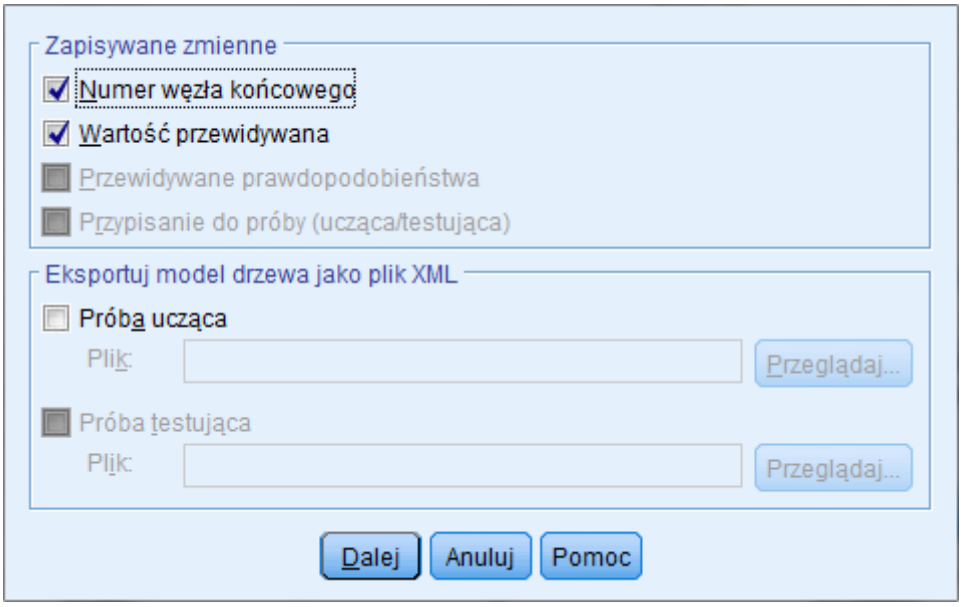


	Producent	Model					Pojemnosc_silnika_dm3	Konie_mechaniczne	Rozstaw_osi_cale
1	Acura	Integra					1.8	140	101.2
2	Acura	TL					3.2	225	108.1
3	Acura	CL					3.2	225	106.9
4	Acura	RL					3.5	210	114.6
5	Audi	A4					1.8	150	102.6
6	Audi	A6					2.8	200	108.7
7	Audi	A8							
8	BMW	323i							
9	BMW	328i							
10	BMW	528i							
11	Buick	Century							
12	Buick	Regal							
13	Buick	Park Avenue							
14	Buick	LeSabre					3.8	205	112.2
15	Cadillac	DeVille					4.6	275	115.3
16	Cadillac	Seville					4.6	275	112.2
17	Cadillac	Eldorado					4.6	275	108.0
18	Cadillac	Catera					3.0	200	107.4
19	Cadillac	Escalade					5.7	255	117.5
20	Chevrolet	Cavalier	145.519	9.250	Passenger	13.260	2.2	115	104.1
21	Chevrolet	Malibu	135.126	11.225	Passenger	16.535	3.1	170	107.0
22	Chevrolet	Lumina	24.629	10.310	Passenger	18.890	3.1	175	107.5
23	Chevrolet	Monte Carlo	42.593	11.525	Passenger	19.390	3.4	180	110.5
24	Chevrolet	Camaro	26.402	13.025	Passenger	24.340	3.8	200	101.1
25	Chevrolet	Corvette	17.947	36.225	Passenger	45.705	5.7	345	104.5
26	Chevrolet	Prizm	32.299	9.125	Passenger	13.960	1.8	120	97.1
27	Chevrolet	Metro	21.855	5.160	Passenger	9.235	1.0	55	93.1
28	Chevrolet	Impala	107.995	.	Passenger	18.890	3.4	180	110.5
29	Chrysler	Sebring Coupe	7.854	12.360	Passenger	19.840	2.5	163	103.7
30	Chrysler	Sebring Conv.	32.775	14.180	Passenger	24.495	2.5	168	106.0
31	Chrysler	Concorde	31.148	13.725	Passenger	22.245	2.7	200	113.0
32	Chrysler	Cirrus	32.306	12.640	Passenger	16.480	2.0	132	108.0
33	Chrysler	LHS	13.462	17.325	Passenger	28.340	3.5	253	113.0
34	Chrysler	Town & Country	53.480	19.540	Car	.	.	.	.
35	Chrysler	300M	30.696	.	Passenger	29.185	3.5	253	113.0

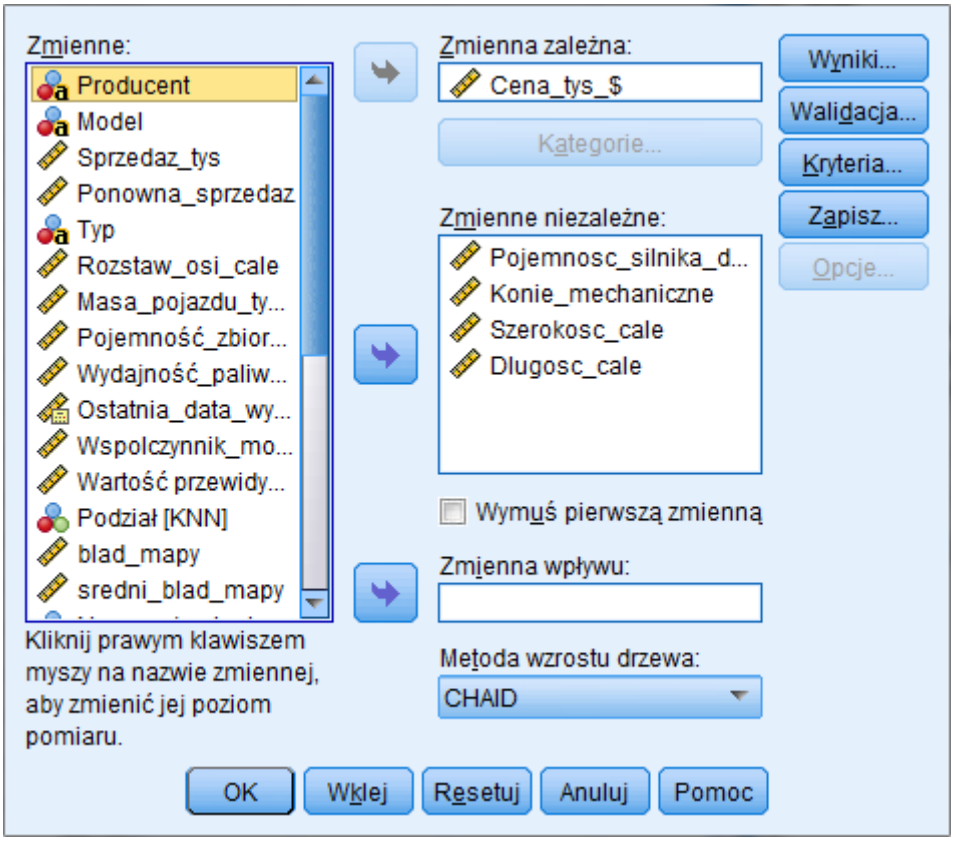
Kryteria


Ograniczenia wzrostu
CRT
Przycinanie
Predyktory substytucyjne
Maksymalna głębokość drzewa
Automatyczna
Maksimum poziomów dla CHAID to 3, zaś dla CRT i QUEST jest równe 5.
Użytkownika
Wartość: 5
Minimalna liczba obserwacji
Węzeł nadrzędny: 20
Węzeł podrzędny: 10
Dalej
Anuluj
Pomoc

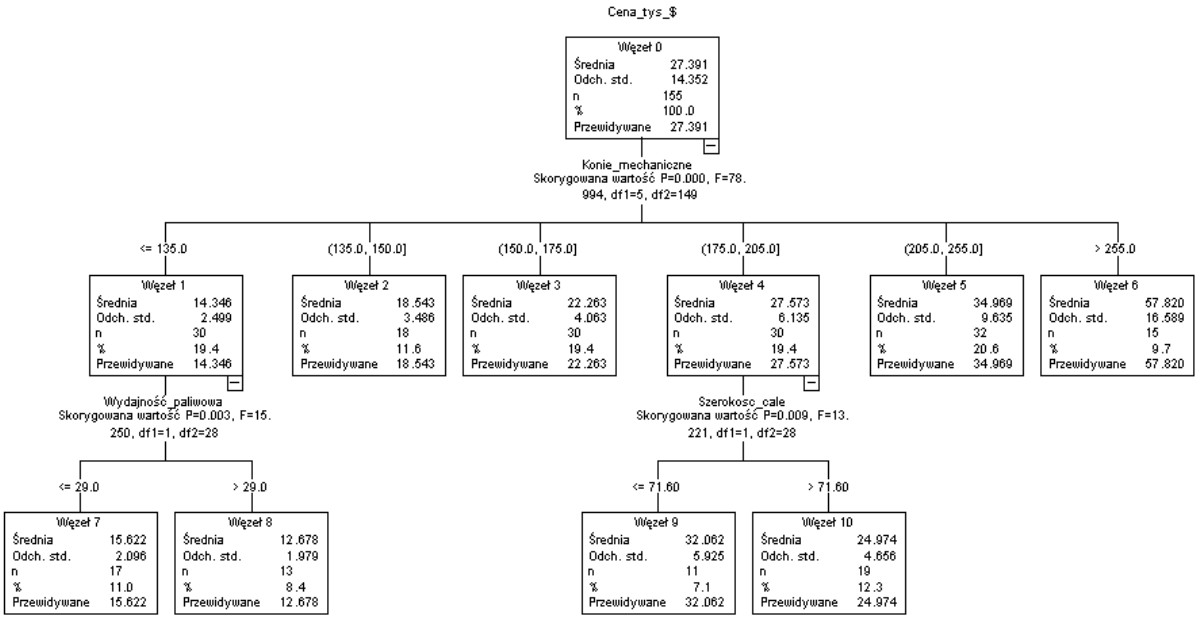
Zapisz


Zapisywane zmienne
Numer węzła końcowego
Wartość przewidywana
Przewidywane prawdopodobieństwa
Przypisanie do próby (ucząca/testująca)
Eksportuj model drzewa jako plik XML
Próba ucząca
Plik:
Przeglądaj...
Próba testująca
Plik:
Przeglądaj...
Dalej
Anuluj
Pomoc

CHAID

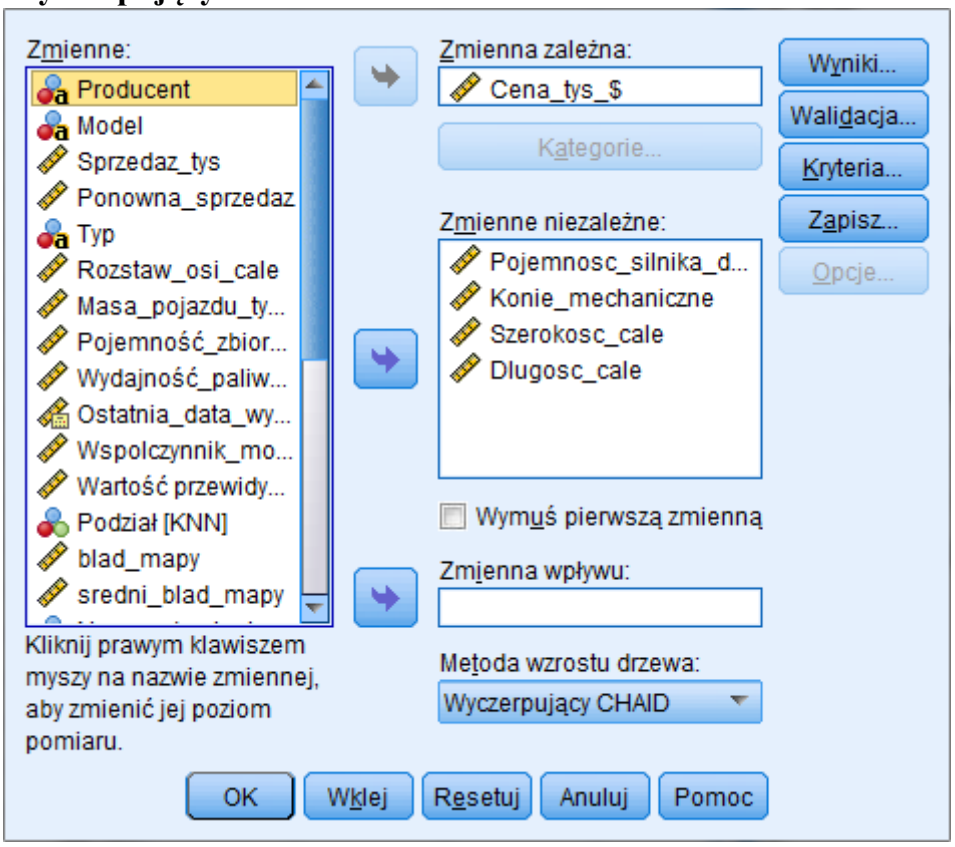


Wyniki	Uwzględnione zmienne niezależne	Konie_mechaniczne, Pojemnosc_silnika_dm3, Szerokosc_cale
	Liczba węzłów	11
	Liczba węzłów końcowych	8
	Głębokość	2

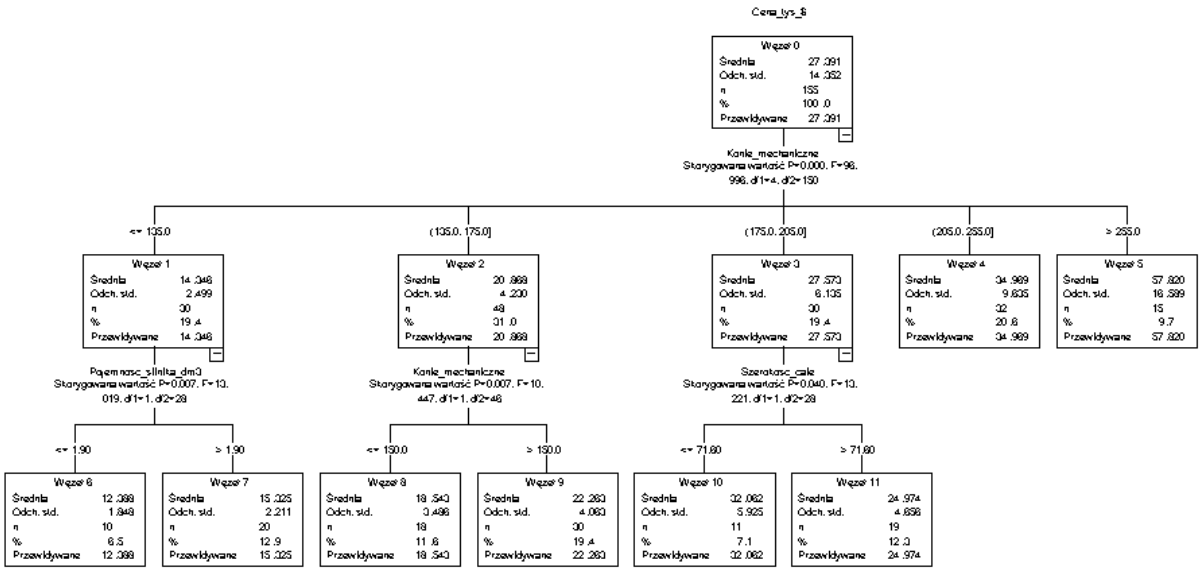


Przy wykorzystaniu metody CHAID do wzrostu drzewa otrzymujemy dosyć szerokie drzewo z 6 węzłami końcowymi. Pierwszym kryterium podziału są Konie_mechaniczne. Przy pierwszym podziale otrzymujemy 2 węzły końcowe – węzeł 5 w którym zawarte są 32 rekordy z wartością w przedziale 205.0-255.0 dla zmiennej Konie_mechanczne oraz węzeł 6 w którym zawarte jest 15 rekordów z wartością w wyższą od 255.0 dla zmiennej Konie_mechanczne. Przewidywana cena w tys $ dla węzła 5 to 34.959, dla węzła 6 wynosi 57.820. Węzeł 1 dzieli się na 2 węzły końcowe. Kryterium podziału jest zmienna Wydajność_paliwowa. Węzeł 7 zawiera 17 rekordów o wydajności mniejszej bądź rownej niż 29 mil na galon, natomiast węzeł 8, 13 rekordów o wydajności większej niż 29 mil na galon. Węzeł 4 dzieli się na 2 węzły końcowe. Kryterium podziału to Szerokość_cale. Węzeł 9 zawiera 11 aut węższych bądź równych 71.60 cala i przewidywana cena wynosi 32.062 tys. $, a Węzeł 10 zawiera 19 samochodów szerszych niż 71.60 cala i przewidywana dla nich cena to 24.974 tys. $.

Wyczerpujący CHAID

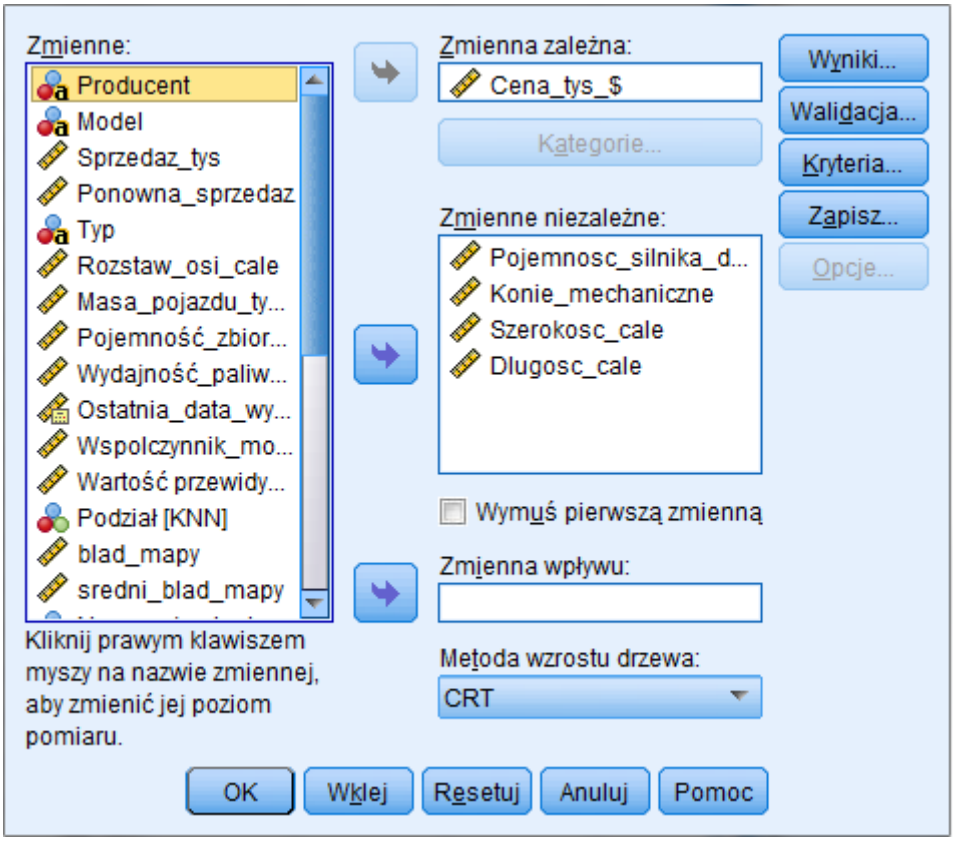


Wyniki	Uwzględnione zmienne niezależne	Konie_mechaniczne, Pojemnosc_silnika_dm3, Szerokosc_cale
	Liczba węzłów	12
	Liczba węzłów końcowych	8
	Głębokość	2

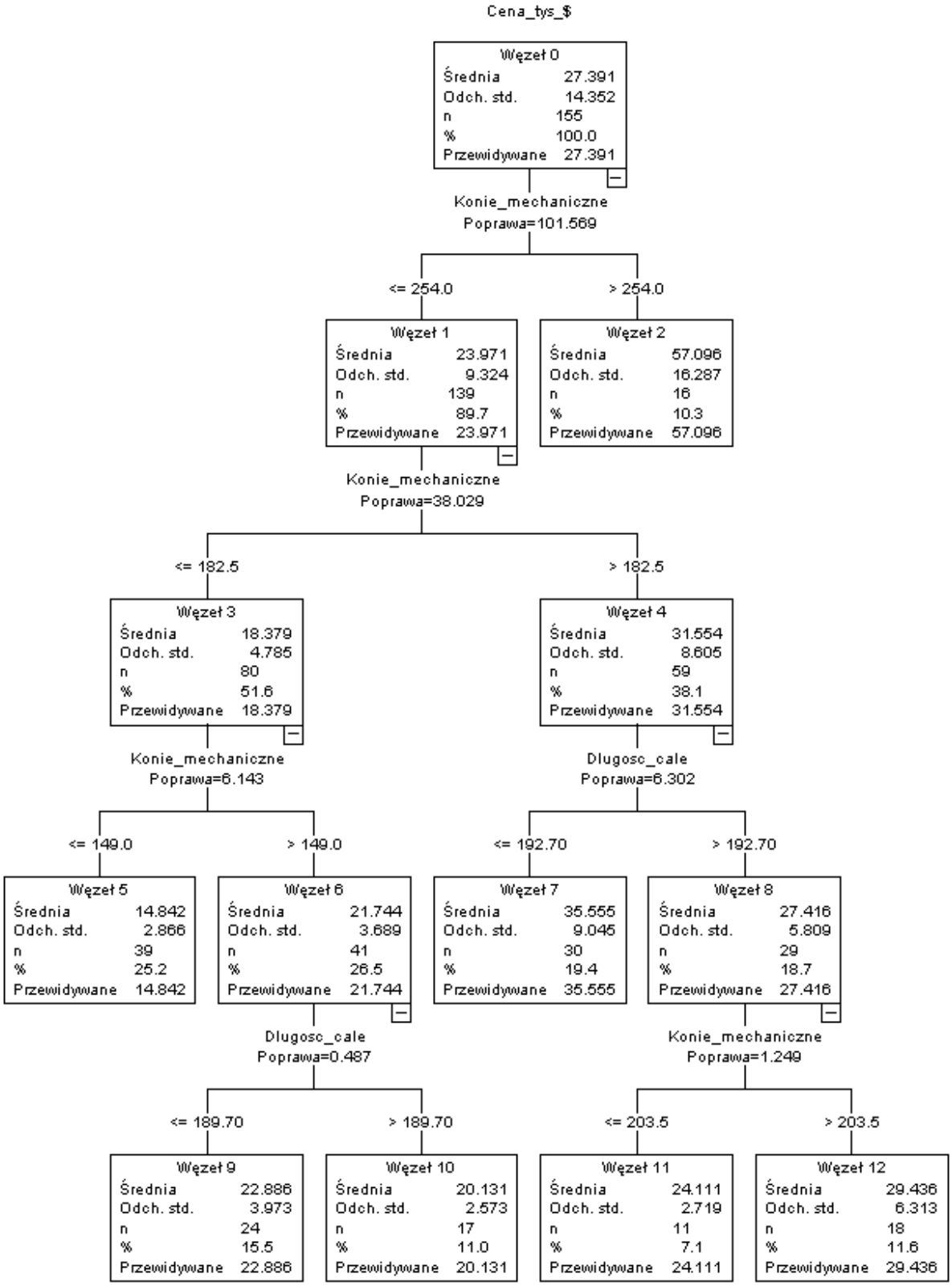


Przy wykorzystaniu metody Wyczerpujący CHAID otrzymujemy 8 węzłów końcowych. Przy pierwszym podziale, dla którego kryterium podziału to Konie_mechaniczne, otrzymujemy 5 węzłów z czego 2 są końcowymi – Węzeł 4 oraz 5. W węźle 4 są auta o mocy w przedziale 205.0-255.0 km i ich przewidywana wartość to 34.969 tys. $. Węzeł 5 zawiera auta o mocy większej niż 255.0km i ich przewidywana wartość to 57.820 tys. $. Każdy z węzłów 1, 2 i 3 dzielą się na 2 węzły końcowe. Węzeł dzieli się na węzeł 6 i 7. Kryterium podziału jest Pojemność_silnika_dm3. Węzeł 6 zawiera pozycje z samochodami i pojemności mniejszej bądź równej 1.9, a węzeł 7 o większej niż 1.9. Wartości przewidywane to odpowiednio 12.388 oraz 15.325 tys. $. Węzeł 8 oraz 9 są węzłami końcowymi odchodzącymi od węzła 2. Kryterium podziału jest takie samo jak pierwsze, tzn. Konie_mechaniczne. Węzeł 8 zawiera auta poniżej bądź równo posiadające 150km, a ich przewidywana cena to 18.543 tys. $. Węzeł 9 natomiast zawiera pojazdy mające ponad 150km, a ich przewidywana cena to 22.263 tys. $. Węzeł 10 oraz 11 są podzielone przez Szerokosc_cale. Węzeł zawiera 10 pojazdów węższych bądź równych 71.60 cala, a ich przewidywana cena to 32.062 tys. $, natomiast węzeł 11 zawiera 19 aut szerszych niż 71.60 cala i ich przewidywany koszt to 24.974 cala.

CRT



Wyniki	Uwzględnione zmienne niezależne	Konie_mechaniczne, Pojemnosc_silnika_dm3, Szerokosc_cale, Dlugosc_cale
	Liczba węzłów	13
	Liczba węzłów końcowych	7
	Głębokość	4


Cena_tys_$
Węzeł 0
Średnia 27.391
Odch. std. 14.352
n 155
% 100.0
Przewidywane 27.391
Konie_mechaniczne
Poprawa=101.569
<= 254.0
> 254.0
Węzeł 1
Średnia 23.971
Odch. std. 9.324
n 139
% 89.7
Przewidywane 23.971
Węzeł 2
Średnia 57.096
Odch. std. 16.287
n 16
% 10.3
Przewidywane 57.096
Konie_mechaniczne
Poprawa=38.029
<= 182.5
> 182.5
Węzeł 3
Średnia 18.379
Odch. std. 4.785
n 80
% 51.6
Przewidywane 18.379
Węzeł 4
Średnia 31.554
Odch. std. 8.605
n 59
% 38.1
Przewidywane 31.554
Konie_mechaniczne
Poprawa=6.143
Długosc_cale
Poprawa=6.302
<= 149.0
> 149.0
<= 192.70
> 192.70
Węzeł 5
Średnia 14.842
Odch. std. 2.866
n 39
% 25.2
Przewidywane 14.842
Węzeł 6
Średnia 21.744
Odch. std. 3.689
n 41
% 26.5
Przewidywane 21.744
Węzeł 7
Średnia 35.555
Odch. std. 9.045
n 30
% 19.4
Przewidywane 35.555
Węzeł 8
Średnia 27.416
Odch. std. 5.809
n 29
% 18.7
Przewidywane 27.416
Długosc_cale
Poprawa=0.487
Konie_mechaniczne
Poprawa=1.249
<= 189.70
> 189.70
<= 203.5
> 203.5
Węzeł 9
Średnia 22.886
Odch. std. 3.973
n 24
% 15.5
Przewidywane 22.886
Węzeł 10
Średnia 20.131
Odch. std. 2.573
n 17
% 11.0
Przewidywane 20.131
Węzeł 11
Średnia 24.111
Odch. std. 2.719
n 11
% 7.1
Przewidywane 24.111
Węzeł 12
Średnia 29.436
Odch. std. 6.313
n 18
% 11.6
Przewidywane 29.436

Przy diagramie drzewa decyzyjnego tworzonym za pomocą metody wzrostu CRT otrzymaliśmy drzewo posiadające 7 węzłów końcowych. Przy pierwszym podziale za pomocą kryterium Konie_mechaniczne dostaliśmy 2 węzły, z czego węzeł 2 jest końcowym. Zawiera on 16 aut o mocy większej niż 254km, a ich przewidywana cena to 57.096 tys. $. Węzeł 1 dzieli się na 2 węzły. Kryterium podziału ponownie są Konie_mechaniczne. Ponieważ Węzeł 1 zawiera wyniki mniejsze bądź równe 254.0, w Węźle 4 są pojazdy o mocy z przedziału od 182.5km do 254 włącznie. Jest ich 59, a ich przewidywana cena to 31.554 tys. $. Dzieli się on następnie na 2 węzły. Węzeł 3, zawierający 80 aut o mocy mniejszej bądź równej 182.5km, dzieli się na 2 węzły, jeden z nich jest końcowym. Kryterium podziału po raz kolejny są Konie_mechaniczne. W Węźle 5 jest 39 pojazdów o mocy mniejszej bądź równej 149.0km, ich przewidywana cena to 14.842 tys. $. Węzeł 6, zawierający 41 samochodów o mocy większej niż 149 km i mniejszej bądź równej 182.5km, dzieli się na 2 węzły, za pomocą kryterium Dlugosc_cale. Węzeł 9 zawiera 24 samochody węższe bądź równe 189.70 cala. Ich przewidywana cena to 22.886 tys. $. Węzeł 10 zawiera 17 samochodów szerszych niż 189.70 cala. Ich przewidywana cena to 20.131 tys. $. Węzeł 4 dzieli się na 2 węzły, kryterium podziału to Dlugosc_cale. Węzeł 7 zawiera 30 pojazdów, węższych bądź równych 192.70 cala, a ich przewidywana cena to 35.555 tys. $. Węzeł 8 natomiast zawiera 29 samochodów, szerszych niż 192.70 cala, a ich przewidywana cena to 27.416 tys. $. Dzieli się on na 2 węzły końcowe, przy pomocy kryterium Konie_mechaniczne. Węzeł 11 zawiera auta o mocy mniejszej, bądź równej 203.5km. Przewidywana ich cena to 24.111 tys. $. W Węźle 12 mamy natomiast 18 samochodów o mocy większej niż 203.5km, a ich przewidywana cena to 29.436 tys. $.

Ryzyko

Ryzyko

Ocena	Błąd standardowy
53.424	8.949

Algorytm budowy drzewa: Wyczerpujący CHAID
Zmienna zależna: Cena_tys_$

Ryzyko

Ocena	Błąd standardowy
53.424	8.949

Algorytm budowy drzewa: Wyczerpujący CHAID
Zmienna zależna: Cena_tys_$

Ryzyko

Ocena	Błąd standardowy
50.863	9.083

Algorytm budowy drzewa: CRT
Zmienna zależna: Cena_tys_$

Wykorzystując Wyczerpujący CHAID oraz CHAID jako metodę wzrostu drzewa, otrzymaliśmy najmniejszy błąd standardowy.

	wartość przewidywana	wartość zależna	
8	18.543	21.5	15.95%
4	34.969	28.4	18.79%
4	34.969	42	20.11%
8	18.543	23.99	29.37%
11	24.974	33.95	35.94%
5	57.82	62	7.23%
9	22.263	26.99	21.23%

Wykorzystując wartość przewidywaną oraz wartość zależną Cena_tys_$, średnia różnica wynosi 16,74%.

Porównanie metod.

Metoda wykres pudełko-wąsy jest bardzo przejrzysta, ponieważ widać gołym okiem zależności przy niektórych zmiennych, niestety porównuje ona tylko 2 zmienne jednocześnie. Przy metodzie k-średnich możemy uwzględnić więcej zmiennych naraz.

Wykorzystując metoda drzewa klasyfikacyjnego otrzymaliśmy większy błąd niż przy wykorzystaniu metody k-najbliższego sąsiada. Przy drzewie klasyfikacyjnym wartość błędu wynosi 16,74%, a przy metodzie k-najbliższego sąsiada w przybliżeniu 12,32%.

Naszym zdaniem, najlepszą metodą do analizy tej bazy danych okazała się metoda k-najbliższego sąsiada. Jest prosta i przejrzysta w użyciu, a co najważniejsze otrzymaliśmy przy niej najmniejszy błąd więc jest również najbardziej precyzyjna w swoich kalkulacjach.