

Przykład eksploracji danych

Case 1.X

JERZY STEFANOWSKI



TPD – Zaawansowana eksploracja danych
edycja 2009/2010

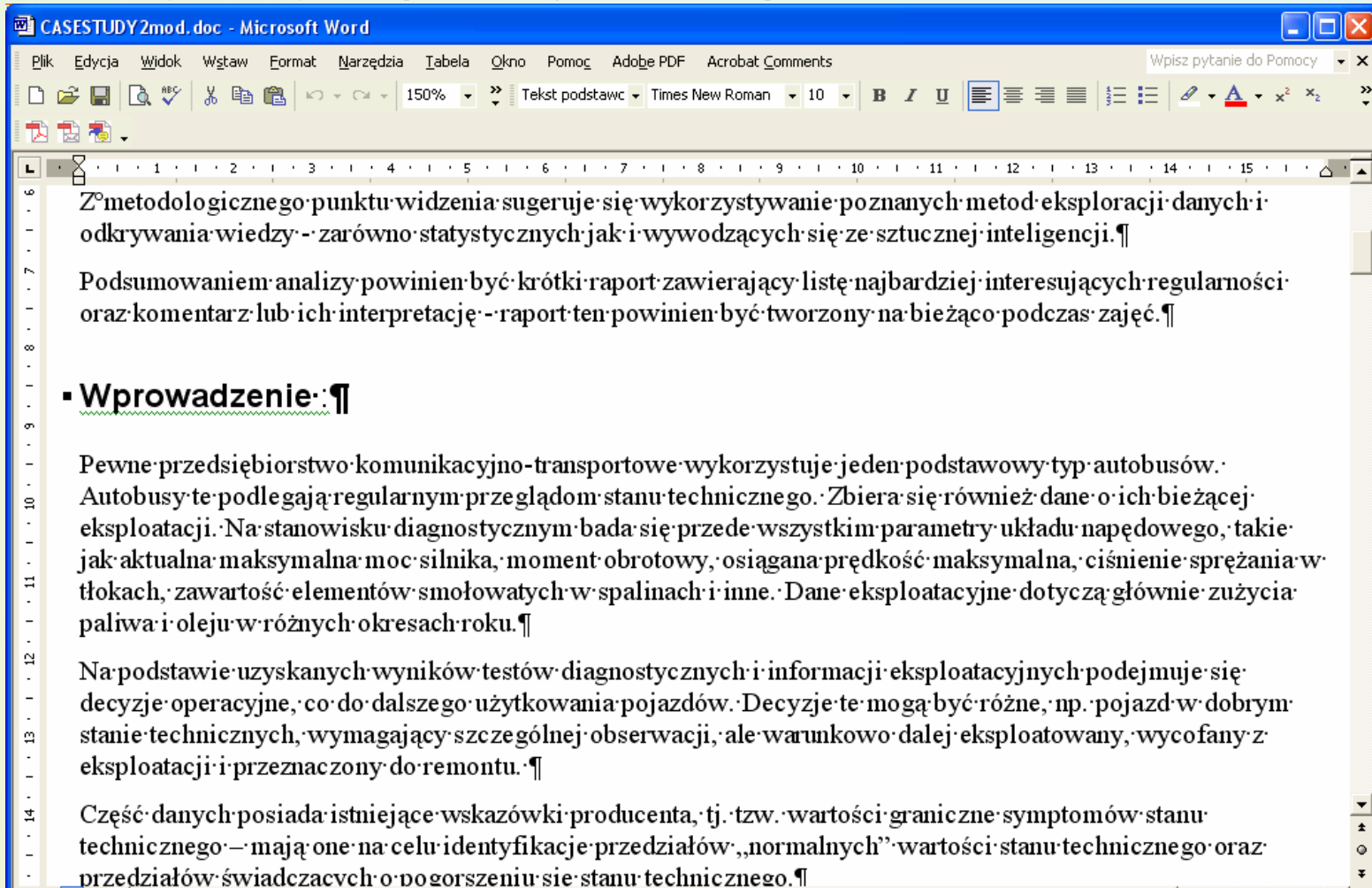
Plan

1. Przykładowe studium przypadku
2. Analiza opisu przypadku
3. Ustalenie celu analizy i scenariusza postępowania
4. Ocena poprawności danych - „czyszczenie”
5. Badanie jakości danych (współzależność)
6. Ocena ważności atrybutów
7. Odkrywanie wiedzy klasyfikacyjnej
 1. Różne podejścia
 2. Ocena zdolności klasyfikacyjnej
 3. Możliwości interpretacji wiedzy
8. Wymagania do sprawozdania



Analiza diagnostycznej bazy danych

- ❑ Problem dotyczy analizy stanu technicznego autobusów używanych przez jedno z przedsiębiorstw w Polsce.



Z metodologicznego punktu widzenia sugeruje się wykorzystywanie poznanych metod eksploracji danych i odkrywania wiedzy - zarówno statystycznych jak i wywodzących się ze sztucznej inteligencji. ¶

Podsumowaniem analizy powinien być krótki raport zawierający listę najbardziej interesujących regularności oraz komentarz lub ich interpretację - raport ten powinien być tworzony na bieżąco podczas zajęć. ¶

▪ Wprowadzenie: ¶

Pewne przedsiębiorstwo komunikacyjno-transportowe wykorzystuje jeden podstawowy typ autobusów. Autobusy te podlegają regularnym przeglądom stanu technicznego. Zbiera się również dane o ich bieżącej eksploatacji. Na stanowisku diagnostycznym bada się przede wszystkim parametry układu napędowego, takie jak aktualna maksymalna moc silnika, moment obrotowy, osiągnięta prędkość maksymalna, ciśnienie sprężania w tłokach, zawartość elementów smołowych w spalinach i inne. Dane eksploatacyjne dotyczą głównie zużycia paliwa i oleju w różnych okresach roku. ¶

Na podstawie uzyskanych wyników testów diagnostycznych i informacji eksploatacyjnych podejmuje się decyzje operacyjne, co do dalszego użytkowania pojazdów. Decyzje te mogą być różne, np. pojazd w dobrym stanie technicznym, wymagający szczególnej obserwacji, ale warunkowo dalej eksploatowany, wycofany z eksploatacji i przeznaczony do remontu. ¶

Część danych posiada istniejące wskazówki producenta, tj. tzw. wartości graniczne symptomów stanu technicznego - mają one na celu identyfikację przedziałów „normalnych” wartości stanu technicznego oraz przedziałów świadczących o pogorszeniu się stanu technicznego. ¶

Analiza diagnostycznej bazy danych

- Bada się stan techniczny 80 autobusów tego samego typu (dokładnie ich silników) na podstawie symptomów stanu technicznego - parametrów pochodzących z okresowych badań diagnostycznych
 - Pierwsza klasyfikacja D1: autobusy są podzielone na dwie klasy: dobry i zły stan techniczny pojazdu
 - Możliwa jest druga klasyfikacja D2 + stan przejściowy
- Cel analizy
 - Ocenia się jakość diagnostyczną symptomów stanu technicznego
 - (pośrednio ocena przydatności tzw. wartości granicznych)
 - Ocena ważności poszczególnych symptomów
 - Ewentualność rankingu lub selekcji
 - Poszukuje się zależności pomiędzy wartościami najistotniejszych w tych symptomów a przydziałem do klas
 - Konstruuje się klasyfikator stanu technicznego

Oryginalny format danych (isf)

```
autobusy.isf - Notatnik
Plik Edycja Format Wzrost Pomoc

**ATTRIBUTES
A0:      (omit) | number of example
MaxSpeed:      (continuous) | km/h
Compr_preature:      (continuous) | Mpa
blacking:      (continuous) | blacking components in exhaust gas [%]
torque:      (continuous) | Nm
summer_cons:      (continuous) | summer fuel consumption l/100km
winter_cons:      (continuous) | winter fuel consumption l/100km
oil_cons:      (continuous) | oil consumption l/1000km
horsepower:      (continuous) | max horesepowewr Km
D1:      [1,2] | technical condtion of a vehicle 1 - good, 2 - bad
D2:      (omit) | [1,2,3]
decision: D1

**EXAMPLES
1, 90, 2.52, 38, 481, 21.8, 26.4, 0.7, 145, 1, 1
2, 76, 2.11, 70, 420, 22.0, 25.5, 2.7, 110, 2, 3
3, 63, 1.96, 82, 400, 22.0, 24.8, 3.7, 101, 2, 3
4, 90, 2.48, 49, 477, 21.9, 25.1, 1.0, 138, 1, 1
5, 85, 2.45, 52, 460, 21.8, 25.2, 1.4, 130, 1, 2
6, 72, 2.20, 73, 425, 23.1, 27.4, 2.8, 112, 2, 3
7, 88, 2.50, 50, 480, 21.6, 24.7, 1.1, 140, 1, 1
8, 87, 2.48, 56, 465, 22.8, 27.6, 1.4, 135, 1, 1
9, 90, 2.56, 16, 486, 26.5, 27.3, 0.2, 150, 1, 1
10, 60, 1.95, 95, 400, 23.3, 24.8, 4.4, 96, 2, 3
11, 80, 2.41, 60, 451, 21.7, 26.1, 1.7, 125, 1, 2
12, 78, 2.4, 63, 448, 21.8, 26.0, 1.9, 120, 1, 2
13, 90, 2.58, 26, 482, 22.4, 24.5, 0.4, 148, 1, 1
14, 62, 1.98, 93, 400, 22, 28.4, 3.9, 100, 2, 3
15, 82, 2.5, 54, 461, 22.0, 26.3, 1.4, 132, 1, 2
```

Analiza nagłówka

- ❑ 80 obserwacji (autobusów)
- ❑ 8 atrybutów
 - max speed km/h
 - Compression pressure Mpa
 - blacking components in exhaust gas [%]
 - torque Nm
 - summer fuel consumption l/100km
 - winter fuel consumption l/100km
 - oil consumption l/1000km
 - max horsepower kW
- ❑ D1: [1,2] | technical conditions of a vehicle: 1 - good, 2 - bad
- ❑ D2: [1,2,3]

Wartości graniczne symptomów

CASESTUDY2mod.doc - Microsoft Word

Plik Edycja Widok Wstaw Format Narzędzia Tabela Okno Pomoc Adobe PDF Acrobat Comments

Wpisz pytanie do Pomocy

150% Tekst podstawowy Times New Roman 10 B I U

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15

▪ Symptomy mające w oryginale wartości liczbowe zostały przetransformowane na wartości dyskretne porządkowe w oparciu o wartości graniczne; Poniżej propozycja pochodząca od eksperta diagnostyki samochodowej:¶

▪ $s_1 \rightarrow : \rightarrow (-\infty, 74>, (74, 79>, (79, 85>, (85, \infty)¶$

▪ $s_2 \rightarrow : \rightarrow (-\infty, 2.2>, (2.2, 2.4>, (2.4, \infty)¶$

▪ $s_3 \rightarrow : \rightarrow (-\infty, 59>, (59, \infty)¶$

▪ $s_4 \rightarrow : \rightarrow (-\infty, 441>, (441, \infty)¶$

▪ $s_5 \rightarrow : \rightarrow (-\infty, 22>, (22, \infty)¶$

▪ $s_6 \rightarrow : \rightarrow (-\infty, 25.2>, (25.2, \infty)¶$

▪ $s_7 \rightarrow : \rightarrow (-\infty, 1.2>, (1.2, \infty)¶$

▪ $s_8 \rightarrow : \rightarrow (-\infty, 119>, (119, \infty)¶$

¶

Jeśli to potrzebne możesz rozważyć powyższą propozycję, ale podstawową analizę proszę wykonać dla danych zdefiniowanych na skalach numerycznych.¶

¶

Ogląd wczytanych danych

STATISTICA - autobusyplain.sta

File Edit View Insert Format Statistics Graphs Tools Data Window Help

Add to Workbook Add to Report

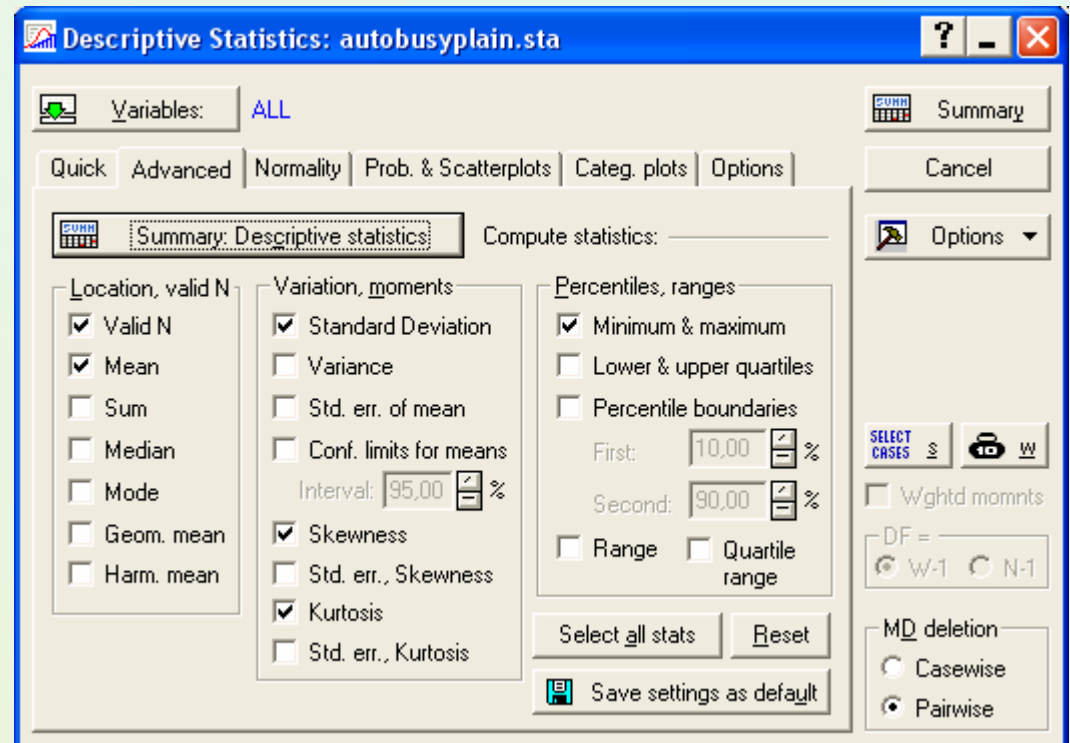
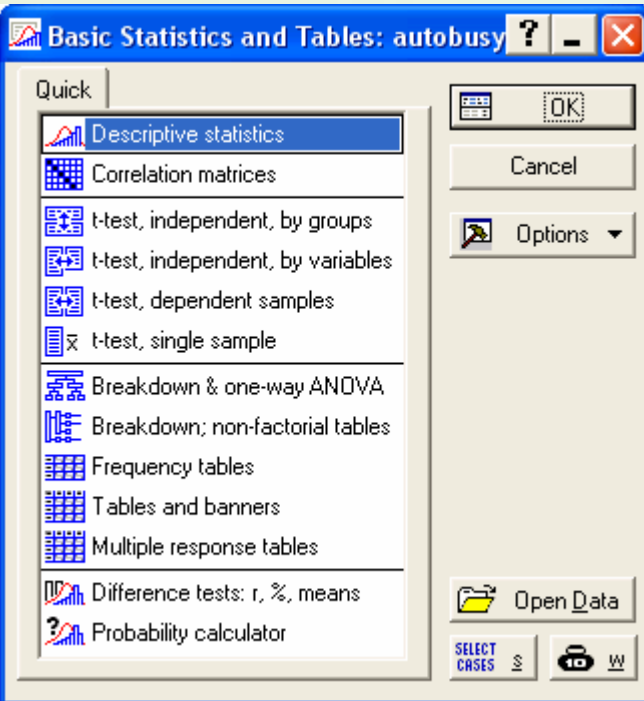
Arial 10 B I U

Data: autobusyplain.sta* (11v by 81c)

	1 id	2 MaxSpeed	3 Compr_preature	4 blacking	5 torque	6 summer_cons	7 winter_cons	8 oil_cons	9 horsepower	10 D1	11 D2
1	1	90	2,52	38	481	21,8	26,4	0,7	145	1	1
2	2	76	2,11	70	420	22	25,5	2,7	110	2	3
3	3	63	1,96	82	400	22	24,8	3,7	101	2	3
4	4	90	2,48	49	477	21,9	25,1	1	138	1	1
5	5	85	2,45	52	460	21,8	25,2	1,4	130	1	2
6	6	72	2,2	73	425	23,1	27,4	2,8	112	2	3
7	7	88	2,5	50	480	21,6	24,7	1,1	140	1	1
8	8	87	2,48	56	465	22,8	27,6	1,4	135	1	1
9	9	90	2,56	16	486	26,5	27,3	0,2	150	1	1
10	10	60	1,95	95	400	23,3	24,8	4,4	96	2	3
11	11	80	2,41	60	451	21,7	26,1	1,7	125	1	2
12	12	78	2,4	63	448	21,8	26	1,9	120	1	2
13	13	90	2,58	26	482	22,4	24,5	0,4	148	1	1
14	14	62	1,98	93	400	22	28,4	3,9	100	2	3
15	15	82	2,5	54	461	22	26,3	1,4	132	1	2
16	16	65	2,22	67	402	22	23,9	2,3	103	2	3
17	17	90	2,48	51	468	22	26,5	1,2	138	1	1
18	18	90	2,6	15	488	20	23,2	0,1	150	1	1
19	19	76	2,39	65	428	27	33,4	2	116	2	3
20	20	85	2,42	50	454	21,5	26,3	1,3	129	1	2
21	21	85	2,41	58	450	22	25,5	1,5	126	1	2
22	22	88	2,47	48	458	22,4	25,1	1,1	130	1	1
23	23	60	1,93	90	400	24	28,7	4	95	2	3
24	24	64	2,2	71	420	23,1	25,2	2,6	105	2	3
25	25	75	2,39	64	432	22,2	25,1	1,7	114	2	2
26	26	74	2,36	64	420	21,9	25,4	1,9	110	2	2
27	27	68	2,15	70	400	22	26	2,6	100	2	3

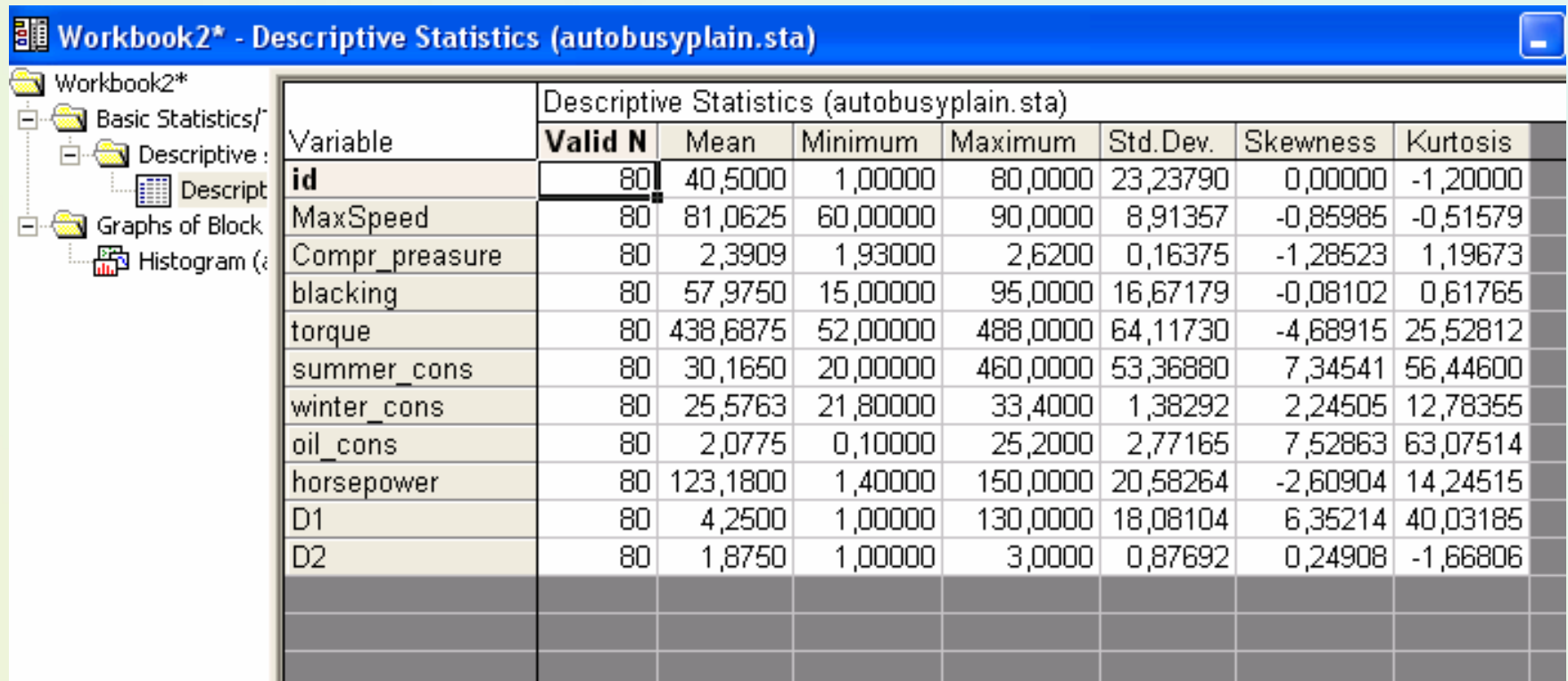
Ocena jakości - statystyki podstawowe

❑ Statistica - descriptive statistics



Raport podstawowych statystyk opisowych

- ❑ Przeanalizujemy podstawowe miary

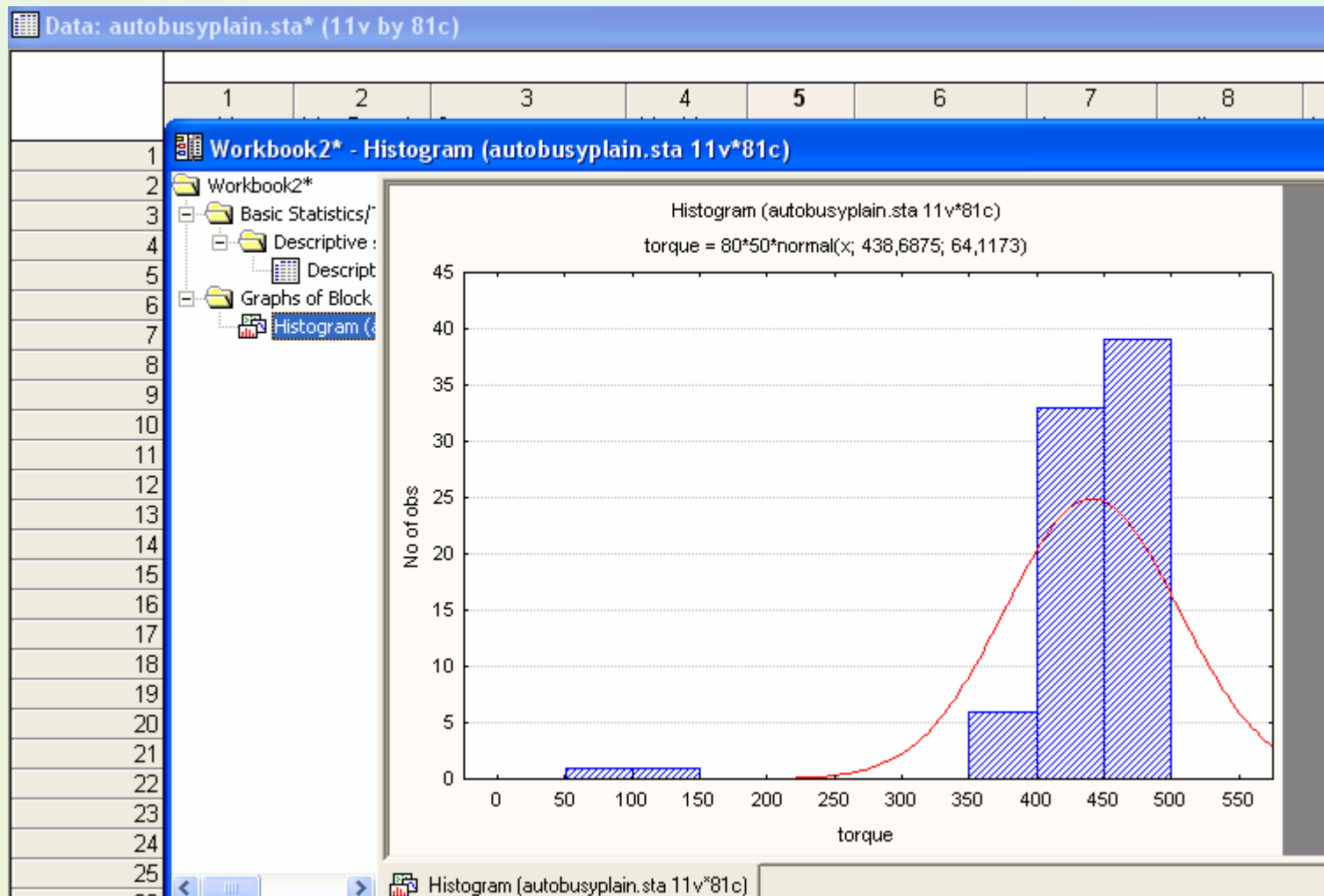


Workbook2* - Descriptive Statistics (autobusyplain.sta)

Variable	Descriptive Statistics (autobusyplain.sta)						
	Valid N	Mean	Minimum	Maximum	Std.Dev.	Skewness	Kurtosis
id	80	40,5000	1,00000	80,0000	23,23790	0,00000	-1,20000
MaxSpeed	80	81,0625	60,00000	90,0000	8,91357	-0,85985	-0,51579
Compr_preasure	80	2,3909	1,93000	2,6200	0,16375	-1,28523	1,19673
blacking	80	57,9750	15,00000	95,0000	16,67179	-0,08102	0,61765
torque	80	438,6875	52,00000	488,0000	64,11730	-4,68915	25,52812
summer_cons	80	30,1650	20,00000	460,0000	53,36880	7,34541	56,44600
winter_cons	80	25,5763	21,80000	33,4000	1,38292	2,24505	12,78355
oil_cons	80	2,0775	0,10000	25,2000	2,77165	7,52863	63,07514
horsepower	80	123,1800	1,40000	150,0000	20,58264	-2,60904	14,24515
D1	80	4,2500	1,00000	130,0000	18,08104	6,35214	40,03185
D2	80	1,8750	1,00000	3,0000	0,87692	0,24908	-1,66806

Spójrzmy dokładniej na wybrane atrybuty

- Np. moment obrotowy (porównaj z definicją dziedziny)



Poszukiwanie źródła trudności

STATISTICA - autobusypain.sta

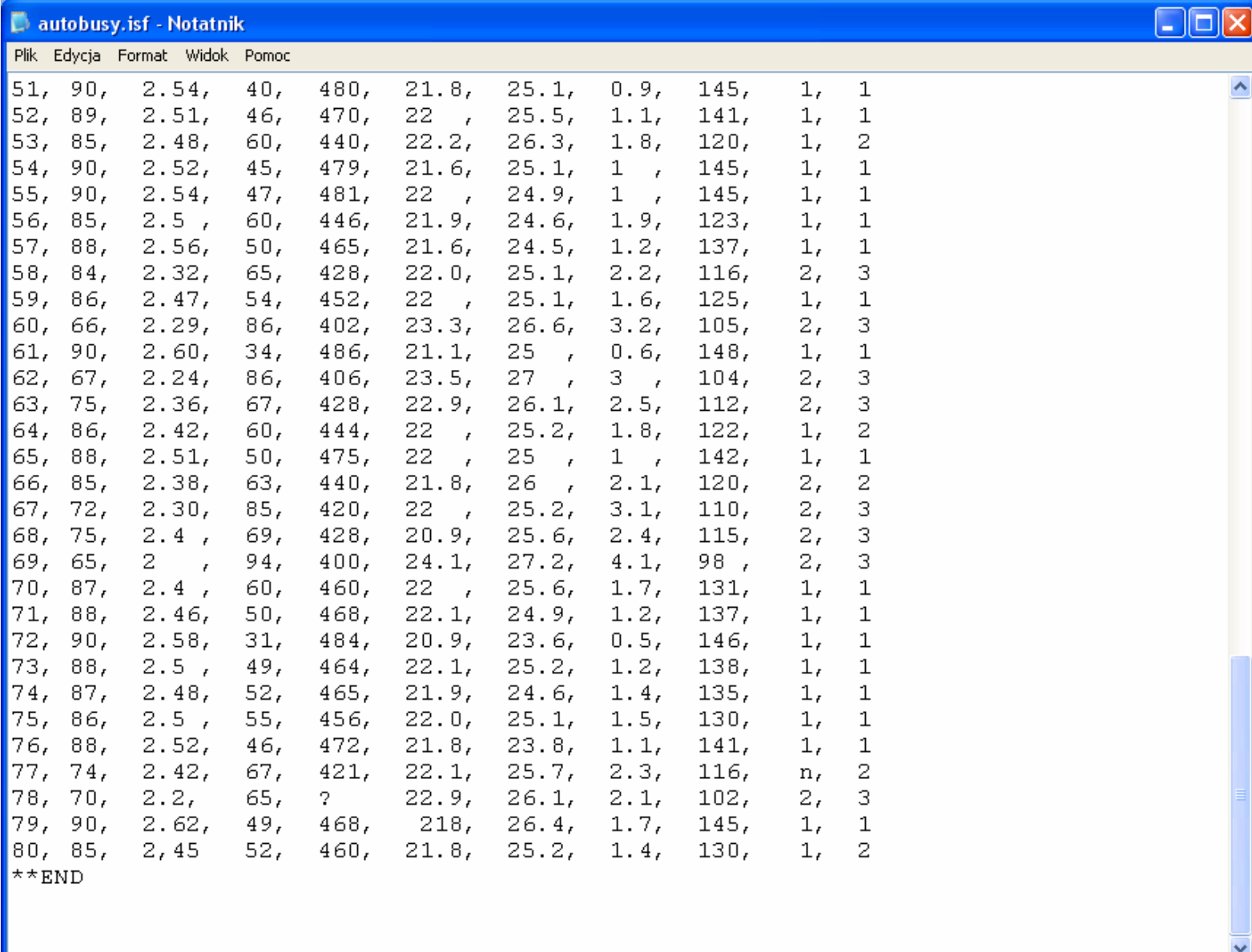
File Edit View Insert Format Statistics Graphs Tools Data Window Help

10 Arial B I U

Data: autobusypain.sta* (11v by 81c)

	1	2	3	4	5	6	7	8	9	10	11
	id	MaxSpeed	Compr. pressure	blacking	torque	summer_cons	winter_cons	oil_cons	horsepower	D1	D2
61	61	90	2,6	34	486	21,1	25	0,6	148	1	1
62	62	67	2,24	86	406	23,5	27	3	104	2	3
63	63	75	2,36	67	428	22,9	26,1	2,5	112	2	3
64	64	86	2,42	60	444	22	25,2	1,8	122	1	2
65	65	88	2,51	50	475	22	25	1	142	1	1
66	66	85	2,38	63	440	21,8	26	2,1	120	2	2
67	67	72	2,3	85	420	22	25,2	3,1	110	2	3
68	68	75	2,4	69	428	20,9	25,6	2,4	115	2	3
69	69	65	2	94	400	24,1	27,2	4,1	98	2	3
70	70	87	2,4	60	460	22	25,6	1,7	131	1	1
71	71	88	2,46	50	468	22,1	24,9	1,2	137	1	1
72	72	90	2,58	31	484	20,9	23,6	0,5	146	1	1
73	73	88	2,5	49	464	22,1	25,2	1,2	138	1	1
74	74	87	2,48	52	465	21,9	24,6	1,4	135	1	1
75	75	86	2,5	55	456	22	25,1	1,5	130	1	1
76	76	88	2,52	46	472	21,8	23,8	1,1	141	1	1
77	77	74	2,42	67	421	22,1	25,7	2,3	116	n	2
78	78	70	2,2	65 ?		22,9	26,1	2,1	102	2	3
79	79	90	2,62	49	468	218	26,4	1,7	145	1	1
80	80	85	2	45	52	460	21,8	25,2	1,4	130	1
81											

Spójrz do oryginalnego pliku



autobusy.isf - Notatnik

Plik	Edycja	Format	Widok	Pomoc						
51,	90,	2.54,	40,	480,	21.8,	25.1,	0.9,	145,	1,	1
52,	89,	2.51,	46,	470,	22,	25.5,	1.1,	141,	1,	1
53,	85,	2.48,	60,	440,	22.2,	26.3,	1.8,	120,	1,	2
54,	90,	2.52,	45,	479,	21.6,	25.1,	1,	145,	1,	1
55,	90,	2.54,	47,	481,	22,	24.9,	1,	145,	1,	1
56,	85,	2.5,	60,	446,	21.9,	24.6,	1.9,	123,	1,	1
57,	88,	2.56,	50,	465,	21.6,	24.5,	1.2,	137,	1,	1
58,	84,	2.32,	65,	428,	22.0,	25.1,	2.2,	116,	2,	3
59,	86,	2.47,	54,	452,	22,	25.1,	1.6,	125,	1,	1
60,	66,	2.29,	86,	402,	23.3,	26.6,	3.2,	105,	2,	3
61,	90,	2.60,	34,	486,	21.1,	25,	0.6,	148,	1,	1
62,	67,	2.24,	86,	406,	23.5,	27,	3,	104,	2,	3
63,	75,	2.36,	67,	428,	22.9,	26.1,	2.5,	112,	2,	3
64,	86,	2.42,	60,	444,	22,	25.2,	1.8,	122,	1,	2
65,	88,	2.51,	50,	475,	22,	25,	1,	142,	1,	1
66,	85,	2.38,	63,	440,	21.8,	26,	2.1,	120,	2,	2
67,	72,	2.30,	85,	420,	22,	25.2,	3.1,	110,	2,	3
68,	75,	2.4,	69,	428,	20.9,	25.6,	2.4,	115,	2,	3
69,	65,	2,	94,	400,	24.1,	27.2,	4.1,	98,	2,	3
70,	87,	2.4,	60,	460,	22,	25.6,	1.7,	131,	1,	1
71,	88,	2.46,	50,	468,	22.1,	24.9,	1.2,	137,	1,	1
72,	90,	2.58,	31,	484,	20.9,	23.6,	0.5,	146,	1,	1
73,	88,	2.5,	49,	464,	22.1,	25.2,	1.2,	138,	1,	1
74,	87,	2.48,	52,	465,	21.9,	24.6,	1.4,	135,	1,	1
75,	86,	2.5,	55,	456,	22.0,	25.1,	1.5,	130,	1,	1
76,	88,	2.52,	46,	472,	21.8,	23.8,	1.1,	141,	1,	1
77,	74,	2.42,	67,	421,	22.1,	25.7,	2.3,	116,	n,	2
78,	70,	2.2,	65,	?	22.9,	26.1,	2.1,	102,	2,	3
79,	90,	2.62,	49,	468,	218,	26.4,	1.7,	145,	1,	1
80,	85,	2.45,	52,	460,	21.8,	25.2,	1.4,	130,	1,	2

**END

Poprawki w pliku

- ❑ Przesunąć wpis wiersza id. 80
- ❑ zapisy ? - wartości średnie w klasie decyzyjnej
- ❑ Symbol spoza dziedziny „n”
 - pewnie niesprawne autobusy - kod 2

Zmiany statystyk opisowych

Workbook2* - Descriptive Statistics (autobusyplainpopraw.sta)

Statistics/Tables (autobusyplainpopraw.sta) Descriptive Statistics dialog Descriptive Statistics (autobusyplainpopraw.sta) of Block Data (autobusyplainpopraw.sta) program (autobusyplainpopraw.sta)

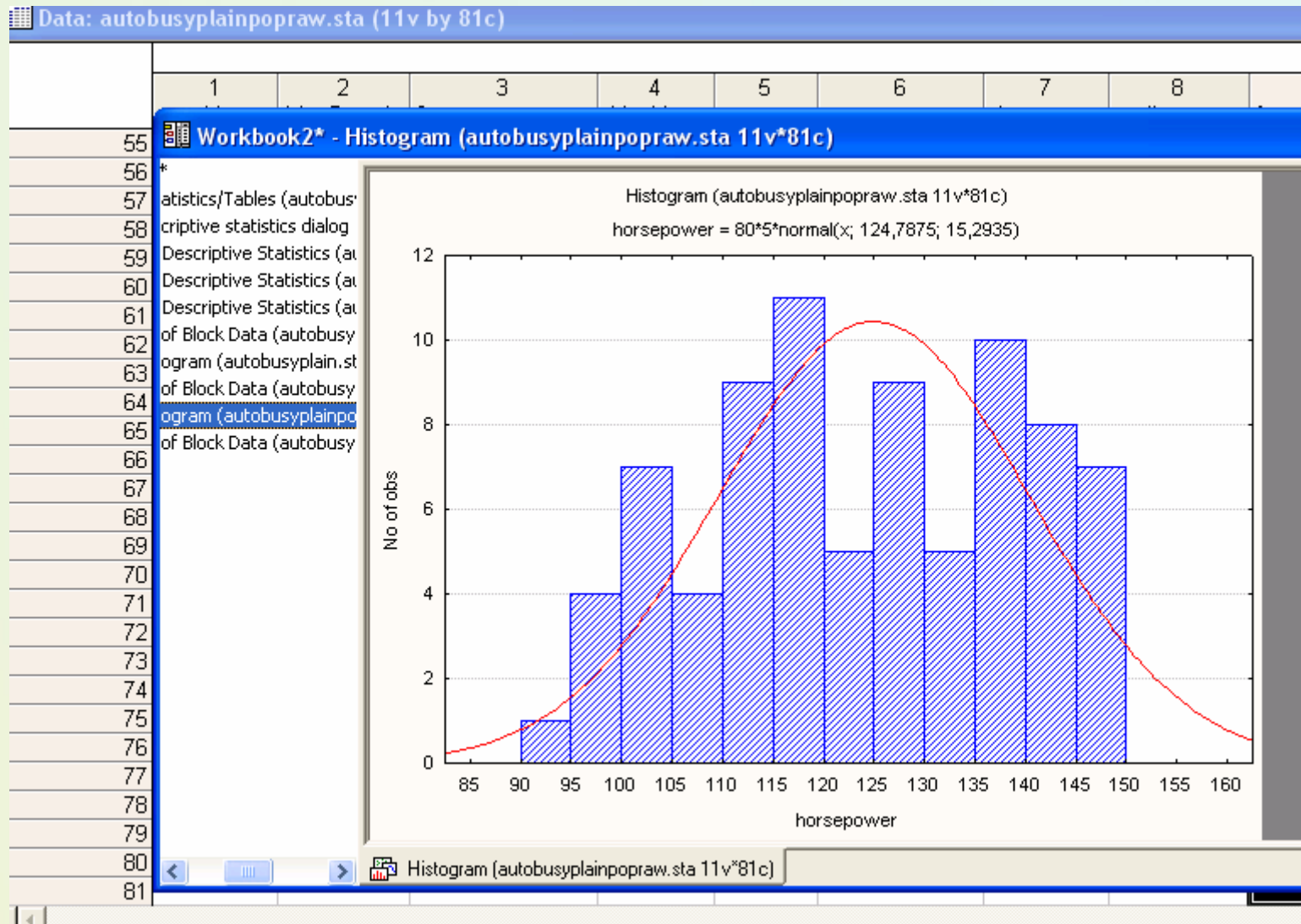
Variable	Descriptive Statistics (autobusyplainpopraw.sta)						
	Valid N	Mean	Minimum	Maximum	Std.Dev.	Skewness	Kurtosis
id	80	40,5000	1,0000	80,0000	23,23790	0,00000	-1,20000
MaxSpeed	80	81,0625	60,0000	90,0000	8,91357	-0,85985	-0,51579
Compr_preature	80	2,3965	1,9300	2,6200	0,15777	-1,34854	1,56840
blackning	80	58,0375	15,0000	95,0000	16,63186	-0,08844	0,65062
torque	80	447,9334	400,0000	488,0000	26,35268	-0,31996	-1,00724
summer_cons	80	24,6875	20,0000	218,0000	21,90940	8,91586	79,65439
winter_cons	80	25,6187	23,2000	33,4000	1,31603	2,81213	14,95687
oil_cons	80	1,7800	0,1000	4,4000	0,91131	0,76994	0,60478
horsepower	80	124,7875	95,0000	150,0000	15,29349	-0,12852	-1,07744
D1	80	1,4000	1,0000	2,0000	0,49299	0,41609	-1,87438
D2	80	1,8875	1,0000	3,0000	0,87140	0,22289	-1,65965

Descriptive Statistics (autobusyplain.sta) Descriptive Statistics (autobusyplainpopraw.sta)

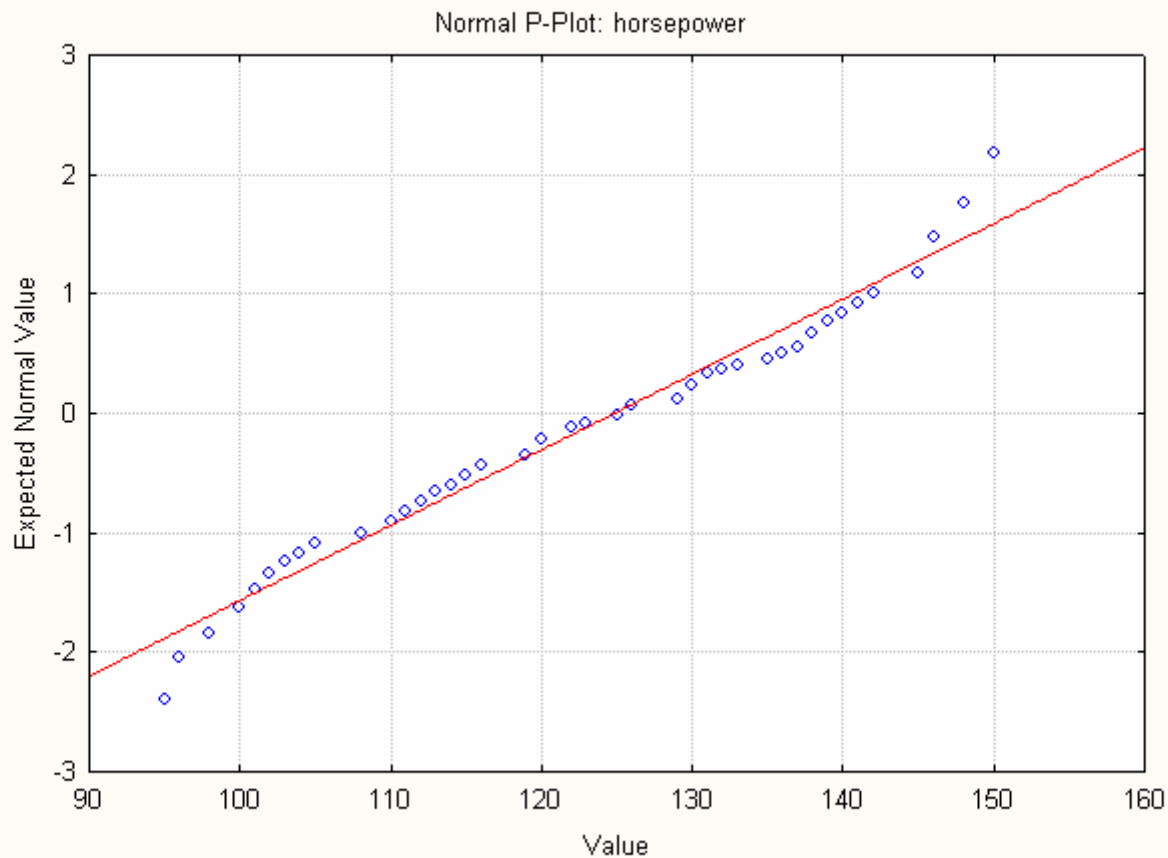
❑ 218? Problem jest z separatorem „ , ”

❑ 218 → 21.8

Sprawdzenie mocy silnika (horsepower)



Spojrzenie na rozkłady



Descriptive Statistics (autobusyplainpopraw.sta)

Descriptive Statistics (autobusyplainpopraw.sta)

Normal P-Plot: horsepower

Analiza współzależności

❑ Macierz korelacji

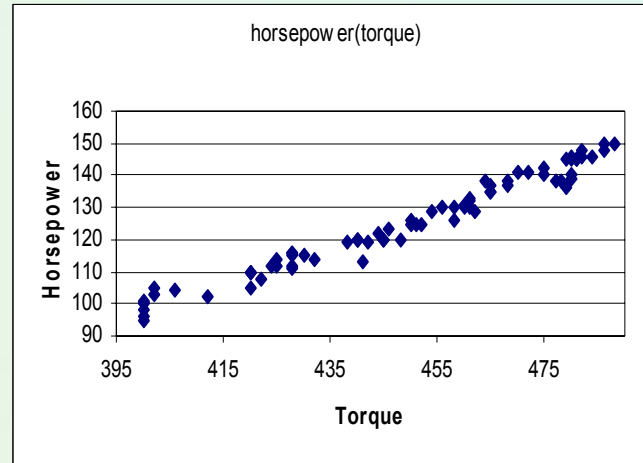
Correlations (autobusyplainpopraw.sta)

Marked correlations are significant at $p < ,05000$
N=80 (Casewise deletion of missing data)

Variable	MaxSpeed	Compr_pressure	blacking	torque	summer_cons	winter_cons	oil_cons	horsepower	D1	D2
MaxSpeed	1,00	0,90	-0,85	0,93	-0,37	-0,34	-0,89	0,93	-0,84	-0,86
Compr_pressure	0,90	1,00	-0,85	0,88	-0,36	-0,34	-0,91	0,89	-0,75	-0,80
blacking	-0,85	-0,85	1,00	-0,90	0,33	0,40	0,96	-0,91	0,73	0,80
torque	0,93	0,88	-0,90	1,00	-0,37	-0,37	-0,92	0,98	-0,85	-0,89
summer_cons	-0,37	-0,36	0,33	-0,37	1,00	0,74	0,35	-0,37	0,38	0,39
winter_cons	-0,34	-0,34	0,40	-0,37	0,74	1,00	0,38	-0,36	0,34	0,40
oil_cons	-0,89	-0,91	0,96	-0,92	0,35	0,38	1,00	-0,91	0,77	0,82
horsepower	0,93	0,89	-0,91	0,98	-0,37	-0,36	-0,91	1,00	-0,84	-0,90

Powiązanie pewnych symptomów

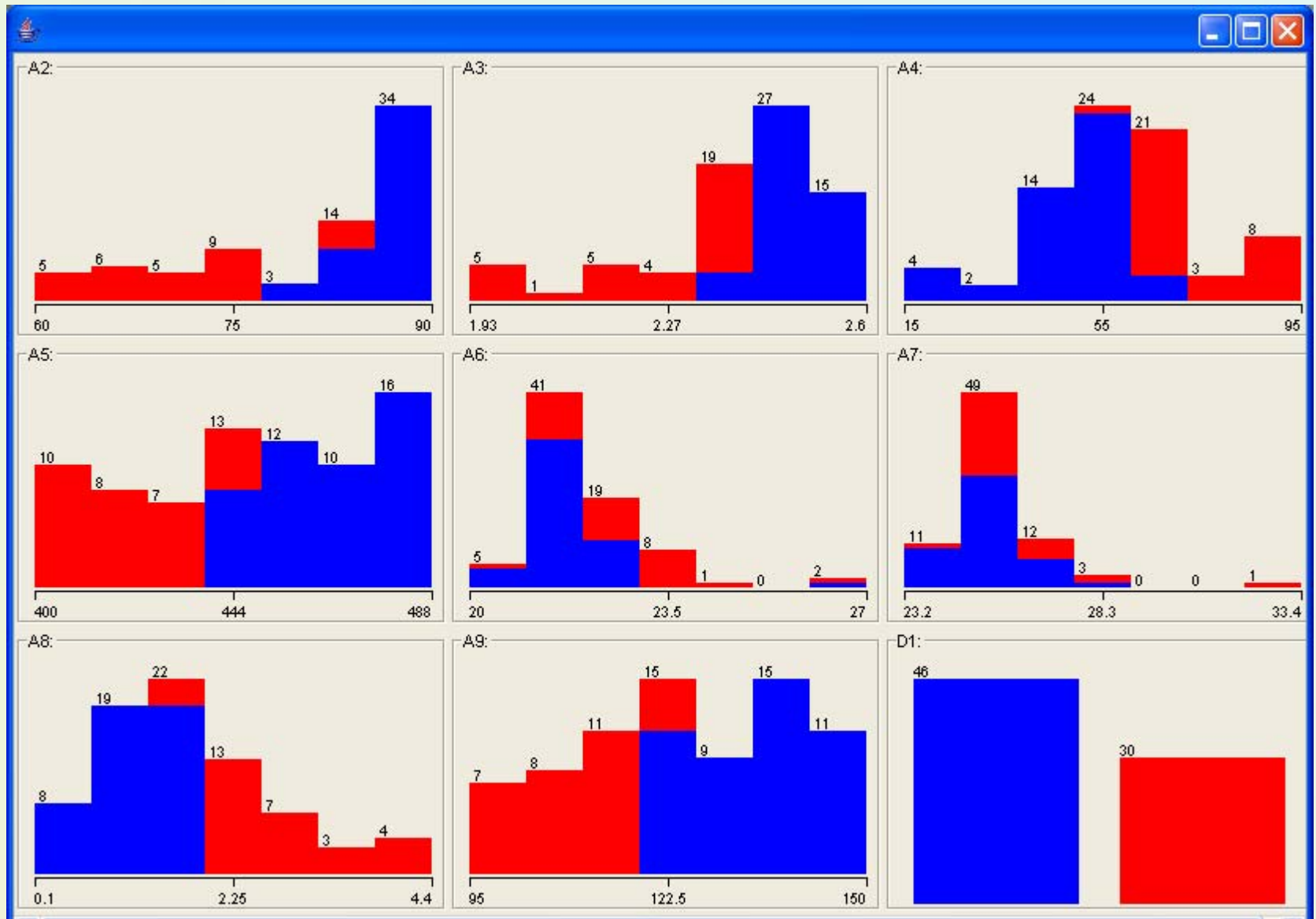
- ❑ Można zrobić wykresy korelacyjne (rozzutu XY)



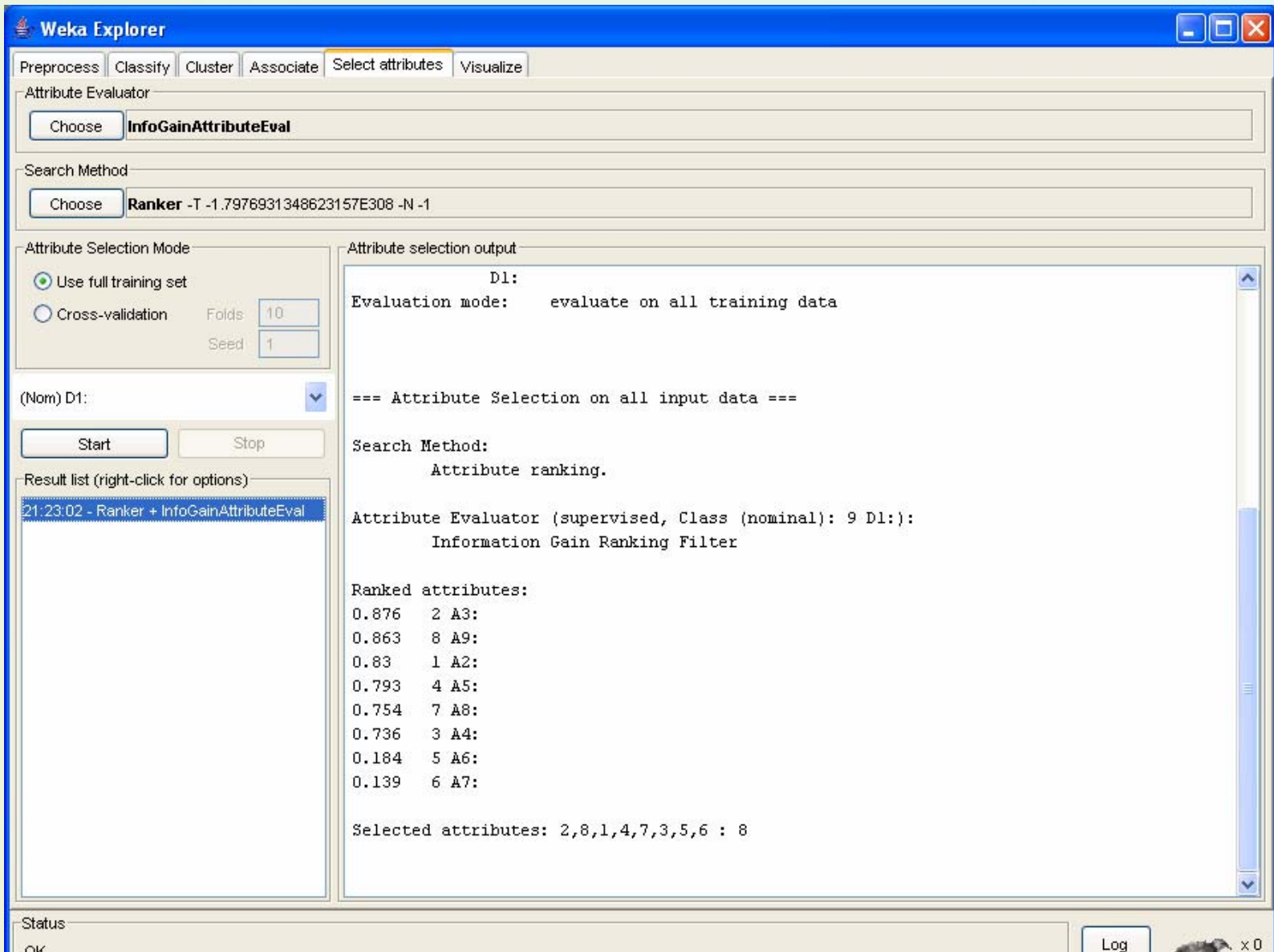
- ❑ Wiedza dziedzinowa

- ciśnienie sprężania powiązane jest z momentem obrotowym, im większe ciśnienie tym większy można uzyskać moment obrotowy,
- moment obrotowy powiązany jest z mocą pojazdu,
- zawartość elementów smółwatych i zużycie oleju świadczyć może o wieku silnika i jego stanie technicznym,
- Mniejsza użyteczność info. o zużyciu paliwa (warunki, styl jazdy,..)

WEKA -visualize



Select attributes z WEKA



Weka Explorer

Preprocess | Classify | Cluster | Associate | **Select attributes** | Visualize

Attribute Evaluator
Choose **InfoGainAttributeEval**

Search Method
Choose **Ranker -T -1.7976931348623157E308 -N -1**

Attribute Selection Mode
 Use full training set
 Cross-validation Folds: 10 Seed: 1

(Nom) D1: ▼

Start Stop

Result list (right-click for options)
21:23:02 - Ranker + InfoGainAttributeEval

Attribute selection output

```
D1:
Evaluation mode:   evaluate on all training data

=== Attribute Selection on all input data ===

Search Method:
  Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 9 D1:):
  Information Gain Ranking Filter

Ranked attributes:
0.876  2 A3:
0.863  8 A9:
0.83   1 A2:
0.793  4 A5:
0.754  7 A8:
0.736  3 A4:
0.184  5 A6:
0.139  6 A7:

Selected attributes: 2,8,1,4,7,3,5,6 : 8
```

Status
OK Log x 0

Relief (ważność z wagowaniem k-NN)

The screenshot shows the Weka Explorer interface with the 'Select attributes' tab active. The 'Attribute Evaluator' is set to 'ReliefAttributeEval -M -1 -D 1 -K 3 -A 2'. The 'Search Method' is 'Ranker -T -1.7976931348623157E308 -N -1'. The 'Attribute Selection Mode' is 'Use full training set' with 10 folds and seed 1. The 'Result list' shows the selected method. The 'Attribute selection output' window displays the following text:

```
=== Attribute Selection on all input data ===  
  
Search Method:  
    Attribute ranking.  
  
Attribute Evaluator (supervised, Class (nominal): 9 D1:):  
    Relief Ranking Filter  
    Instances sampled: all  
    Number of nearest neighbours (k): 3  
    Equal influence nearest neighbours  
  
Ranked attributes:  
0.2458  4 A5:  
0.216   8 A9:  
0.1776  1 A2:  
0.1638  7 A8:  
0.1577  2 A3:  
0.1322  3 A4:  
0.0137  5 A6:  
0.0115  6 A7:  
  
Selected attributes: 4,8,1,7,2,3,5,6 : 8
```

The 'Status' bar at the bottom shows 'OK' and a 'Log' button.

Statistica - Data Miner

The screenshot displays the Statistica software interface. The main window title is "STATISTICA - autobusypainpopraw.sta". The menu bar includes File, Edit, View, Insert, Format, Statistics, Graphs, Tools, Data, and Window. The toolbar contains icons for file operations and a "Resume..." button. The main data table is titled "Data: autobusypainpop" and contains columns for "id" and numerical values. A "Data Miner" menu is open, listing various analysis procedures. A secondary window titled "Cases" is visible on the right, showing a table with columns "power", "10 D1", and "11 D2".

Main Data Table:

	1 id		
	55	55	
	56	56	
	57	57	
	58	58	
	59	59	
	60	60	
	61	61	
	62	62	
	63	63	
	64	64	
	65	65	
	66	66	
	67	67	
	68	68	
	69	69	
	70	70	87
	71	71	88
	72	72	90
	73	73	88
	74	74	87
	75	75	86
	76	76	88
	77	77	74
	78	78	70
	79	79	90
	80	80	85
	81		

Data Miner - My Procedures:

- Data Miner - All Procedures
- Data Miner - Data Cleaning and Filtering
- Data Miner - General Slicer/Dicer Explorer with Drill-Down
- Data Miner - General Classifier (Trees and Clusters)
- Data Miner - General Modeler and Multivariate Explorer
- Data Miner - General Forecaster
- Data Miner - General Neural Network Explorer
- Neural Networks
- Independent Components Analysis
- Generalized EM & k-Means Cluster Analysis
- Association Rules
- Sequence, Association, and Link Analysis
- General Classification/Regression Tree Models
- General CHAID Models
- Interactive Trees (C&RT, CHAID)
- Boosted Tree Classifiers and Regression
- Random Forests for Regression and Classification
- Generalized Additive Models
- MARSplines (Multivariate Adaptive Regression Splines)
- Machine Learning (Bayesian, Support Vectors, Nearest Neighbor)
- Rapid Deployment of Predictive Models (PMML)
- Goodness of Fit, Classification, Prediction
- Feature Selection and Variable Screening
- Combining Groups (Classes) for Predictive Data-Mining

Cases Table:

power	10 D1	11 D2
145	1	1
123	1	1
137	1	1
116	2	3
125	1	1
105	2	3
148	1	1
104	2	3
112	2	3
122	1	2
142	1	1
120	2	2
110	2	3
115	2	3
98	2	3
131	1	1
137	1	1
146	1	1
138	1	1
135	1	1
130	1	1
141	1	1
116	2	2
102	2	3
145	1	1
130	1	2

CART - tree

STATISTICA - Workbook4* - [Tree graph for D1]

File Edit View Insert Format Statistics Graphs Tools Workbook Window Help

Normal Graph [modi...]

Workbook5* - node 1 - 1

Workbook4* - Tree graph for D1

Tree graph for D1
 Num. of non-terminal nodes: 1, Num. of terminal nodes: 2
 Model: C&RT

```

    graph TD
      Node1["ID=1  
N=80  
1"]
      Node2["ID=2  
N=30  
2"]
      Node3["ID=3  
N=60  
1"]
      Node1 -- "<= 2,395000" --> Node2
      Node1 -- "> 2,395000" --> Node3
  
```

Tree graph for D1

node 1 - 1

	11	D2
1	1	1
1	1	1
1	1	1
2	3	3
1	1	1
2	3	3
2	3	3
1	2	2
1	1	1
2	2	2
2	3	3
2	3	3
1	1	1
1	1	1
1	1	1
1	1	1
1	1	1
1	1	1
1	1	1
2	2	2
2	3	3
1	1	1
1	2	2

Przeglądanie węzłów drzewa (dec 1)

Workbook5* - node 1 - 1

- Workbook5*
 - node 1 - 1
 - node 2 - 2
 - node 3 - 1

Trees C&RT Results: autobusyplai

Summary | Classification | Prediction | Report

tree (grow, prune):

- Grow tree
- Brush tree
- Grow tree & prune
- Remove all branches
- Grow tree 1 level
- Remove 1 level

review tree:

- Tree browser
- Scrollable tree
- Tree graph
- Tree layout

node/branch:

Node id: 1

- Grow branch
- Grow branch 1 level
- Predictor stats
- Remove branch
- Split criterion
- SQL code
- Data
- Histogram of DV
- Select a surrogate
- torque
- Surrogate stats

Save tree | New tree | Close | Options

Node 1

Number of cases in node: 80

selected category: 1

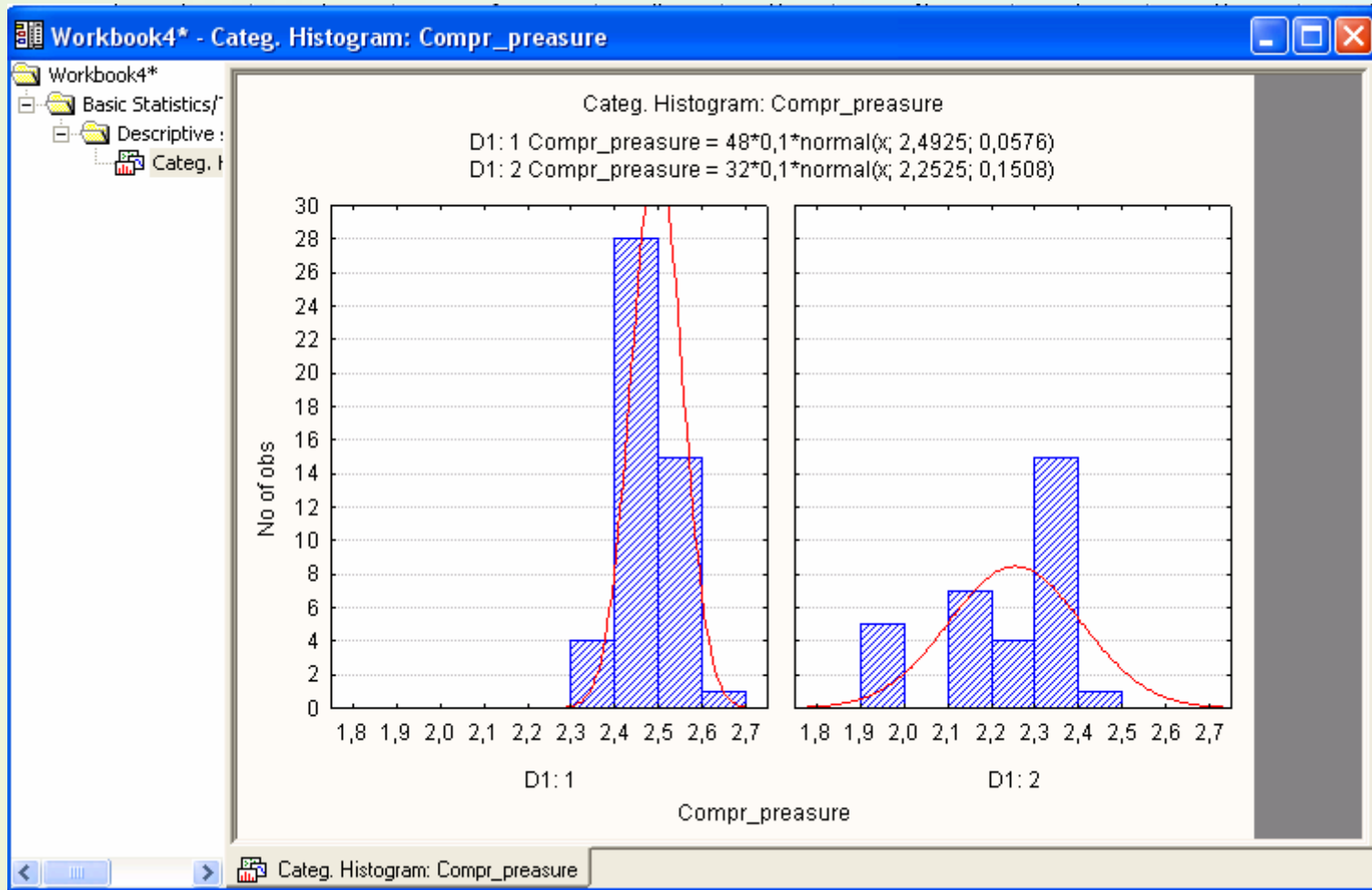
Split Variable: "Compr_preature"

Node 2: $\leq 2,395000$

Node 3: $> 2,395000$

Category	Number of cases
1	48
2	32

Skategoryzowane histogramy



Wykresy histogramowe (skategoryzowane)

STATISTICA - Workbook4* - [Categ. Histogram: horsepower]

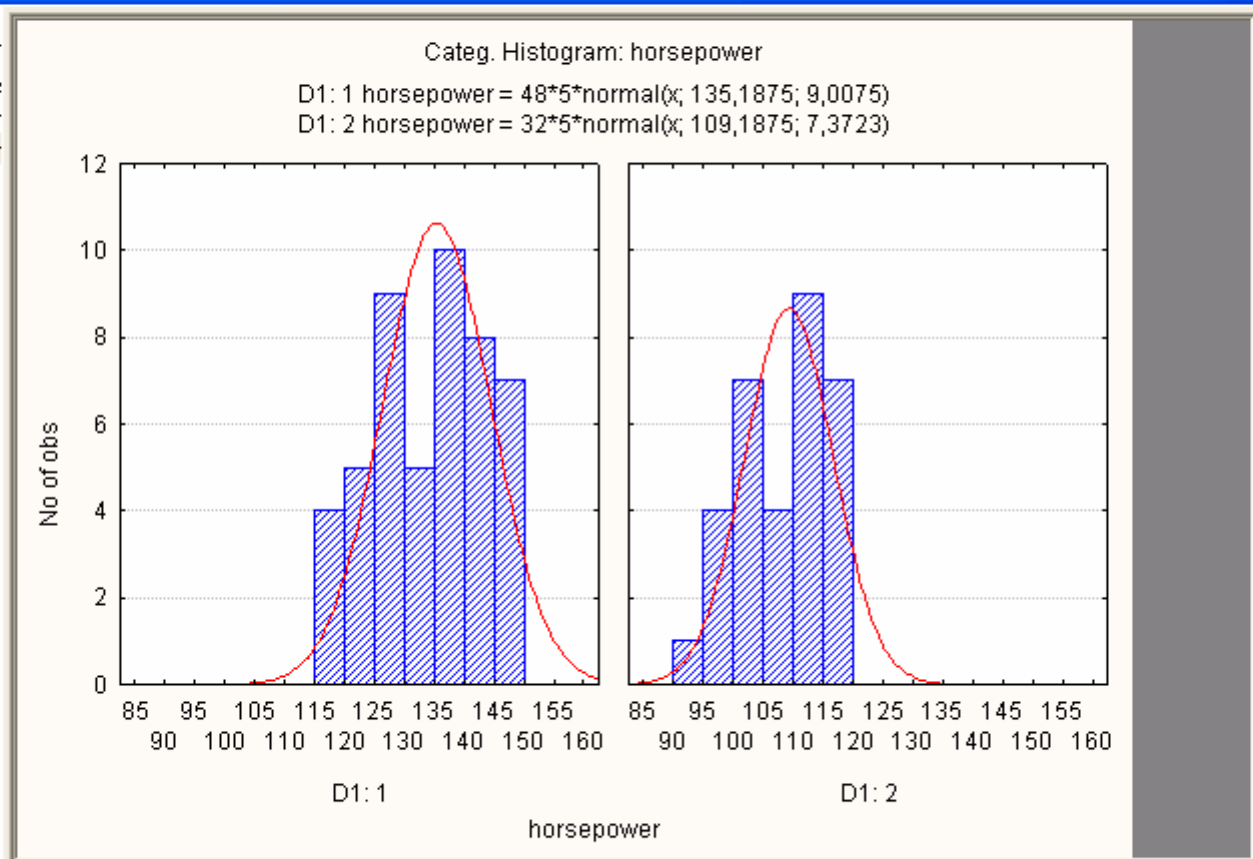
File Edit View Insert Format Statistics Graphs Tools Workbook Window Help

Normal Graph [modi...]

Normal Graph [modi...]

Workbook4* - Categ. Histogram: horsepower

Workbook4*
 Basic Statistics/
 Descriptive:
 Categ. H
 Categ. H



9	10	1
horsepower	D1	D
145	1	
123	1	
137	1	
116	2	
125	1	
105	2	
148	1	
104	2	
112	2	
122	1	
142	1	
120	2	
110	2	
115	2	
98	2	
131	1	
137	1	
146	1	
138	1	
135	1	
130	1	
141	1	
116	2	
102	2	
145	1	

Categ. Histogram: Compr_preature Categ. Histogram: horsepower

79 79 90 2,62 49 468 21,8 26,4 1,7

Podjęcie teorii zbiorów przybliżonych

The image shows the ROSE2 software interface. The main window, titled "ROSE2 - Untitled.ros", has a menu bar (File, View, Project, Method, Tools, Help) and a toolbar with icons for file operations and an "Exit" button. On the left, a project tree lists various methods: Preprocessing, Approximations, Reduction (with sub-items: Core, Lattice Search, Discernibility Matrix, Heuristic Search, Manual Search), Rule Induction, Validation, and Similarity Relation. The main workspace contains a file named "Buseskod.isf" and a large splash screen for ROSE2. The splash screen features a red rose and the text: "ROSE2 Rough Sets Data Explorer", "Version 2.2 (build 2002-11-05)", "© 1999-2002 IDSS", and the URL "http://www-idss.cs.put.poznan.pl/rose".

In the bottom right corner, a "Reduct Viewer" window is open, displaying a table of reducts. The window title is "Reduct Viewer - C:\Usr\Jurek\students\dyp". The table has three columns: "#", "Reduct", and "Length".

#	Reduct	Length
1	Compr_pressure, blacking, torque	3
2	MaxSpeed, oil_cons	2
3	MaxSpeed, Compr_pressure	2
4	Compr_pressure, oil_cons	2
5	Compr_pressure, horsepower	2

Dyskretne dane + teoria zbiorów przybliżonych

Hierarchia ważności symptomów dla klasyfikacji autobusów

- Można ustalać względną ważność symptomów dla klasyfikacji przykładów - przykładowa hierarchia ważności s_8, s_1, s_3, s_4 inne mniej istotne - analiza statystyczna
- w przypadku dyskretyzacji dziedzin wartości w oparciu o normy możliwe ustalenie podzbiorów zredukowanych $\{s_2, s_3, s_8\}, \{s_2, s_3, s_4\}, \{s_1, s_2\}$ - teoria zbiorów przybliżonych
- Notacja s_1 - prędkość maksymalna

Drzewo J4.8 (WEKA)

The screenshot shows the Weka Explorer interface with the 'Classify' tab selected. The classifier chosen is 'J48 -U -M 2'. The 'Test options' section shows 'Cross-validation' selected with 10 folds and 66% split. The 'Result list' shows three entries, with '21:27:16 - trees.J48' selected. The 'Classifier output' pane displays the following text:

```
-----  
A3: <= 2.39: 2 (29.0)  
A3: > 2.39: 1 (47.0/1.0)  
  
Number of Leaves :    2  
  
Size of the tree :    3  
  
Time taken to build model: 0.01 seconds  
  
=== Stratified cross-validation ===  
=== Summary ===  
  
Correctly Classified Instances      75                98.6842 %  
Std Dev. of Corr. Class. Inst.     3.75 %  
Incorrectly Classified Instances    1                1.3158 %  
Kappa statistic                     0.9723  
Mean absolute error                  0.0259  
Root mean squared error              0.116  
Relative absolute error              5.4081 %  
Root relative squared error          23.726 %  
Total Number of Instances           76  
  
=== Detailed Accuracy By Class ===  
  
TP Rate  FP Rate  Precision  Recall  F-Measure  Class  
1        0.033    0.979     1       0.989     1  
0.967    0         1         0.967   0.983     2  
  
=== Confusion Matrix ===
```

C4.5 z trochę innymi parametrami

The screenshot displays the C4.5 software interface for a classification task on the 'autobusy' dataset. The main window shows a comparison of decision trees before and after pruning. A confusion matrix window is open, showing the results for the training set. Three windows at the bottom show the unpruned tree, the pruned tree, and the resulting rules.

C4.5 autobusy (8 attributes, 80 training cases, 0 test cases)

Data Tree Rules Cross-validation Special Help

Tree	Before pruning			After pruning			Estimate
	Size	Errors	Errors (test)	Size	Errors	Errors (test)	
1	5	1 (1.2%)		5	1 (1.2%)		6.0%

Confusion matrix (training set)

Orig. \ C4.5	b	g
b	32	
g	1	47

Unpruned tree

```
Tree
├─ Compr_preasure <= 2.39
│   └─ b
├─ Compr_preasure > 2.39
│   └─ MaxSpeed <= 78
│       └─ b
│       └─ MaxSpeed > 78
│           └─ g
```

Pruned tree

```
Tree
├─ Compr_preasure <= 2.39
│   └─ b
├─ Compr_preasure > 2.39
│   └─ MaxSpeed <= 78
│       └─ b
│       └─ MaxSpeed > 78
│           └─ g
```

Rules

```
Rule 1: [97.1%]
  IF   MaxSpeed > 78
  AND  Compr_preasure > 2.39
  THEN g

Rule 2: [95.5%]
  IF   Compr_preasure <= 2.39
  THEN b

Rule 3: [90.9%]
  IF   MaxSpeed <= 78
  THEN b

Default class: g

Errors in training set: 1 (1.2%)
```

Algorytm MODLEM

The screenshot shows the Weka Explorer application window. The 'Classify' tab is active, and the 'Modlem' classifier is selected. The 'Test options' section shows 'Cross-validation' with 10 folds and 66% split. The 'Classifier output' pane displays the following information:

```
=== Classifier model (full training set) ===  
Rule 1.(A3: >= 2.4)&(A2: >= 75.5) => (D1: = 1); [46, 46, 100%, 100%]  
Rule 2.(A3: < 2.4) => (D1: = 2); [29, 29, 96.67%, 100%]  
Rule 3.(A2: < 75.5) => (D1: = 2); [23, 23, 76.67%, 100%]  
  
Number of rules: 3  
Number of conditions: 4  
  
Time taken to build model: 0.07 seconds  
  
=== Stratified cross-validation ===  
=== Summary ===  
  
Correctly Classified Instances          73           96.0526 %  
Std Dev. of Corr. Class. Inst.         8.966 %  
Incorrectly Classified Instances        3           3.9474 %  
Kappa statistic                        0.9159  
Mean absolute error                    0.0395  
Root mean squared error                0.1987  
Relative absolute error                 8.2469 %  
Root relative squared error            40.6335 %  
Total Number of Instances              76  
  
=== Detailed Accuracy By Class ===  
  
TP Rate  FP Rate  Precision  Recall  F-Measure  Class  
1         0.1      0.939     1       0.968     1  
0.9      0        1         0.9     0.947     2  
  
=== Confusion Matrix ===
```

The 'Result list' on the left shows a single entry: '21.26.02 - rules.Modlem'.

Ripper - alternatywne podejście do indukcji reguł

The screenshot shows the Weka Explorer interface with the 'Classify' tab selected. The classifier is set to 'JRip -F 3 -N 2.0 -O 2 -S 1'. The test options are set to 'Cross-validation' with 10 folds and 66% split. The classifier output shows the following results:

```
D1:
Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

JRIP rules:
=====

(A3: <= 2.39) => D1:=2 (29.0/0.0)
=> D1:=1 (47.0/1.0)

Number of Rules : 2

Time taken to build model: 0.06 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      74           97.3684 %
Std Dev. of Corr. Class. Inst.      5           %
Incorrectly Classified Instances      2           2.6316 %
Kappa statistic                      0.9449
Mean absolute error                   0.0387
```

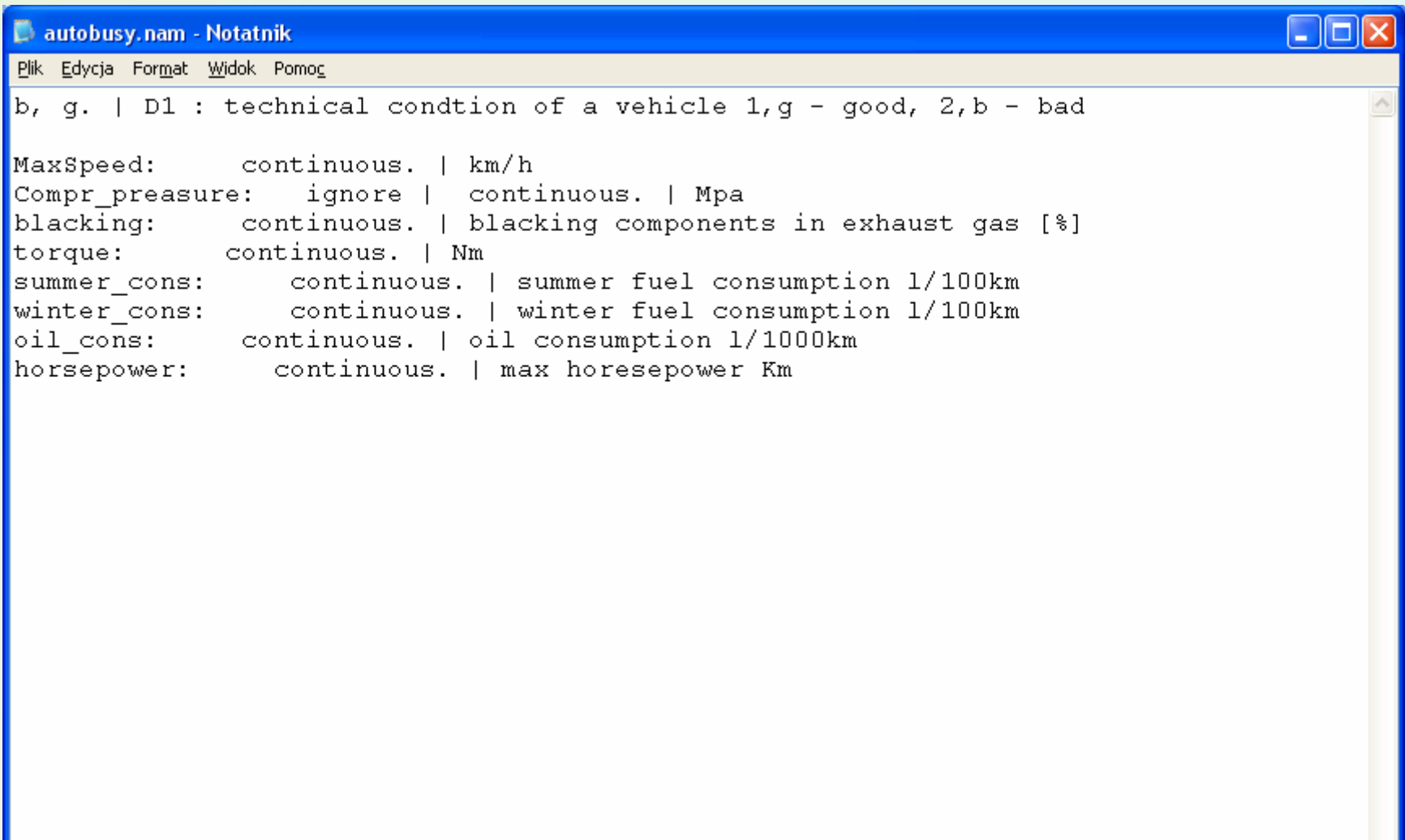
The status bar at the bottom shows 'Status OK' and a 'Log' button.

Podsumowanie klasyfikatorów

- ❑ Dość efektywny klasyfikator
 - Trafności około 97-98%
 - Rozpoznawanie trudniejszej klasy 90%
- ❑ Klasyfikatory symboliczne porównywalne do niesymbolicznych:
 - MLP BP - 97,37%
 - SVM - 97,37% (klasa n 0,93)
 - RBF - 93,27%
 - IBL (3) - 97,37%
 - J4.8 - 98,68%
- ❑ Można dokonać oceny wiedzy symbolicznej
- ❑ Podobną analizę warto przeprowadzić dla drugiej klasyfikacji.

Trudności pomiaru → alternatywne reprezentacje wiedzy klasyfikującej

- ❑ Ciśnienie sprężania → najtrudniejsze do pomiaru



```
autobusy.nam - Notatnik
Plik Edycja Format Widok Pomoc
b, g. | D1 : technical condition of a vehicle 1,g - good, 2,b - bad
MaxSpeed:      continuous. | km/h
Compr_preature: ignore | continuous. | Mpa
blacking:      continuous. | blacking components in exhaust gas [%]
torque:        continuous. | Nm
summer_cons:   continuous. | summer fuel consumption l/100km
winter_cons:   continuous. | winter fuel consumption l/100km
oil_cons:      continuous. | oil consumption l/1000km
horsepower:    continuous. | max horesepower Km
```

Pomijanie trudnych atrybutów (- ciśnienie sprężania)

C4.5 autobusy (8 attributes, 80 training cases, 0 test cases)

Data Tree Rules Cross-validation Special Help

Before pruning				After pruning			
Tree	Size	Errors	Errors (test)	Size	Errors	Errors (test)	Estimate
1	3	2 (2.5%)		3	2 (2.5%)		6.4%

Confusion matrix (training set)

Org. \ C4.5	b	g
b	31	1
g	1	47

Unpruned tree

```
torque <= 441
├── b
└── torque > 441
    └── g
```

Pruned tree

```
torque <= 441
├── b
└── torque > 441
    └── g
```

34

Maskujemy dalsze atrybuty

The screenshot displays a data mining software interface with several windows. The main window, titled "C4.5 autobusy (8 attributes, 80 training cases, 0 test cases)", shows a table comparing decision trees before and after pruning. The table has columns for Tree, Size, Errors, Errors (test), and Estimate. The results show that after pruning, the error rate increases from 0.0% to 0.7%.

Before pruning				After pruning			
Tree	Size	Errors	Errors (test)	Size	Errors	Errors (test)	Estimate
1	5	0 (0.0%)		5	0 (0.0%)		4.8%

The "Unpruned tree" window shows a decision tree with the following structure:

- oil_cons > 1.9
 - b
- oil_cons <= 1.9
 - MaxSpeed <= 76
 - b
 - MaxSpeed > 76
 - g

The "Pruned tree" window shows the same tree structure, but with the node "oil_cons > 1.9" highlighted, indicating it has been pruned.

The "Confusion matrix (training set)" window shows the following matrix:

Org. \ C4.5	b	g
b	32	
g		48

The bottom window, also titled "C4.5 autobusy (8 attributes, 80 training cases, 0 test cases)", shows a table comparing decision trees before and after pruning. The table has columns for Tree, Size, Errors, Errors (test), and Estimate. The results show that after pruning, the error rate increases from 0.0% to 0.7%.

Before pruning				After pruning			
Tree	Size	Errors	Errors (test)	Size	Errors	Errors (test)	Estimate
1	5	1 (1.4%)	0 (0.0%)	5	1 (1.4%)	0 (0.0%)	6.6%
2	3	1 (1.4%)	2 (25.0%)	3	1 (1.4%)	2 (25.0%)	5.5%
3	5	1 (1.4%)	2 (25.0%)	5	1 (1.4%)	2 (25.0%)	6.6%
4	5	0 (0.0%)	0 (0.0%)	5	0 (0.0%)	0 (0.0%)	5.3%
5	5	0 (0.0%)	0 (0.0%)	5	0 (0.0%)	0 (0.0%)	5.3%
6	7	1 (1.4%)	1 (12.5%)	7	1 (1.4%)	1 (12.5%)	8.0%
7	5	0 (0.0%)	0 (0.0%)	5	0 (0.0%)	0 (0.0%)	5.3%
8	5	0 (0.0%)	0 (0.0%)	5	0 (0.0%)	0 (0.0%)	5.3%
9	5	0 (0.0%)	0 (0.0%)	5	0 (0.0%)	0 (0.0%)	5.3%
10	3	1 (1.4%)	2 (25.0%)	3	1 (1.4%)	2 (25.0%)	5.5%
Avg.	4.8	0.5 (0.7%)	0.7 (8.8%)	4.8	0.5 (0.7%)	0.7 (8.8%)	5.9%

Jeszcze inne możliwości

- Dalsze maskowania atrybutów w drzewie prowadzą do niższych trafności
- Jakie jeszcze metody oferują symboliczną reprezentacje wiedzy?
- Jak stworzyć profil/charakterystykę autobusu należącego do określonej klasy stanu technicznego?

Minimalny zbiór reguł klasyfikujących (MODLEM)

1. if ($s_2 \geq 2.4$ MPa) & ($s_7 < 2.1$ l/1000km) then (technical state=good) [46]
2. if ($s_2 < 2.4$ MPa) then (technical state=bad) [29]
3. if ($s_7 \geq 2.1$ l/1000km) then (technical state=bad) [24]

Oszacowana trafność klasyfikowania
(**'leaving one out' test**) 98.7%.

Explore algorithm (Stefanowski, Vanderpooten)

- ❑ Inny cel poszukiwania reguł
 - to extract from data set inducing **all rules** that *satisfy* some *user's requirements* connected with *his interest* (regarding, e.g. the strength of the rule, level of confidence, length, sometimes also emphasis on the syntax of rules).
- ❑ Special technique of exploration the space of possible rules:
 - Progressively generation rules of increasing size using in the most efficient way some 'good' pruning and stopping condition that reject unnecessary candidates for rules.
- ❑ Similar to adaptations of Apriori principle for looking frequent itemsets [AIS94]; Brute [Etzioni]
- ❑ Więcej: J.Stefanowski, D.Vanderpooten: Induction of decision rules in classification and discovery-oriented perspectives, *International Journal of Intelligent Systems*, vol. 16 no. 1, 2001, 13-28.
- ❑ Lub monografia J.Stefanowski: Algorytmy indukcji reguł decyzyjnych w odkrywaniu wiedzy => patrz <http://www.cs.put.poznan.pl/jstefanowski/jspspdf.html>

Poszukiwanie zbioru reguł silnych (EXPLORE)

Próg satysfakcji (51%):

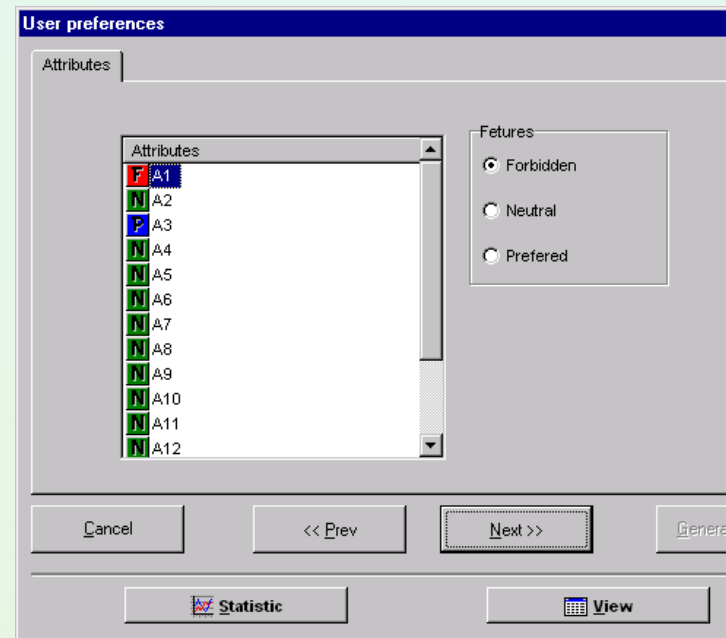
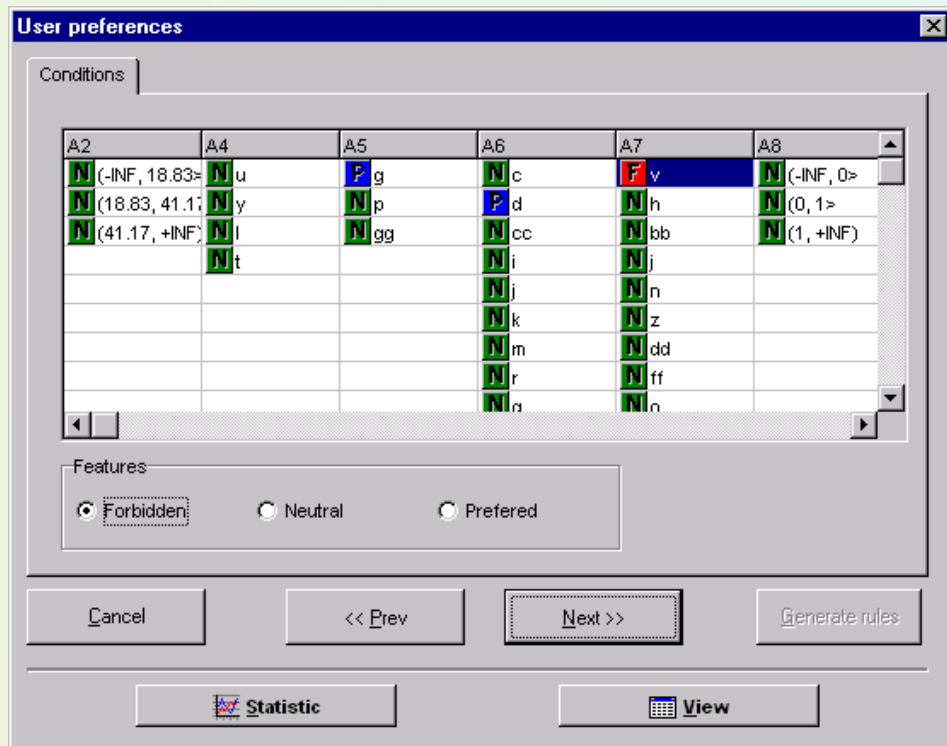
1. if ($s1 > 85$ km/h) then (technical state=good) [34]
2. if ($s8 > 134$ kM) then (technical state=good) [26]
3. if ($s2 \geq 2.4$ MPa) & ($s3 < 61$ %) then (technical state=good) [44]
4. if ($s2 \geq 2.4$ MPa) & ($s4 > 444$ Nm) then (technical state=good) [44]
5. if ($s2 \geq 2.4$ MPa) & ($s7 < 2.1$ //1000km) then (technical state=good) [46]
6. if ($s3 < 61$ %) & ($s4 > 444$ Nm) then (technical state=good) [42]
7. if ($s1 \leq 77$ km/h) then (technical state=bad) [25]
8. if ($s2 < 2.4$ MPa) then (technical state=bad) [29]
9. if ($s7 \geq 2.1$ //1000km) then (technical state=bad) [24]
10. if ($s3 \geq 61$ %) & ($s4 \leq 444$ Nm) then (technical state=bad) [28]
11. if ($s3 \geq 61$ %) & ($s8 < 120$ kM) then (technical state=bad) [27]

Możliwość dodatkowego ograniczania

- ❑ Interactive Explore
- ❑ Sterowanie procedurą poszukiwania reguł decyzyjnych za pomocą następujących parametrów :
 - ustalenie atrybutu decyzyjnego i jego klas decyzyjnych,
 - zbiór preferowanych i zakazanych atrybutów warunkowych,
 - zbiór preferowanych i zakazanych warunków elementarnych,
 - preferowany warunek złożony (część przesłanki),
 - zbiór zakazanych warunków złożonych,
 - maksymalna wielkość zbioru znalezionych reguł,
 - minimalne wsparcie reguły,
 - maksymalna długość reguły,
 - minimalny stopień dyskryminacji (dokładność) reguły.

Sterowanie wyborem

❑ Interactive Explores



Przeglądanie zbioru reguł

RuleEdit - Buseskod.D1.rlf

File Edit View Window Help

Buseskod.D1.rlf: Full View

rule ID	C laxSpeer	C npr_preas	C blacking	C torque	C mmer_co	C nter_coi	C oil_cons	C sepo.	D D1	Length	Cover	Abs. Strength	Rel. Strength	Discr. Level	Generality	Accurad
1								= 2	= 1	01	42	42	0.91	1	0.55	
2	= 1	= 1							= 1	02	46	46	1	1	0.61	
3	= 1		= 0						= 1	02	44	44	0.96	1	0.58	
4	= 1						= 0		= 1	02	46	46	1	1	0.61	
5		= 1	= 0						= 1	02	44	44	0.96	1	0.58	
6		= 1		= 1					= 1	02	45	45	0.98	1	0.59	
7		= 1				= 0			= 1	02	31	31	0.67	1	0.41	
8		= 1					= 0		= 1	02	46	46	1	1	0.61	
9			= 0	= 1					= 1	02	43	43	0.93	1	0.57	
10			= 0			= 0			= 1	02	31	31	0.67	1	0.41	
11				= 1		= 0			= 1	02	31	31	0.67	1	0.41	
12				= 1			= 0		= 1	02	45	45	0.98	1	0.59	
13	= 0								= 2	01	25	25	0.83	1	0.33	
14		= 0							= 2	01	29	29	0.97	1	0.38	
15							= 1		= 2	01	27	27	0.9	1	0.36	
16								= 0	= 2	01	26	26	0.87	1	0.34	
17			= 1	= 0					= 2	02	28	28	0.93	1	0.37	
18			= 1		= 1				= 2	02	15	15	0.5	1	0.2	
19			= 1			= 0			= 2	02	07	07	0.23	1	0.092	
20				= 0	= 1				= 2	02	15	15	0.5	1	0.2	
21				= 0		= 0			= 2	02	07	07	0.23	1	0.092	

```

x Parsing ISF file...
WARNING: Line 4 - truncating 1.5 to integer value
Parsing OK.
File D:\Usr\dyplomanci2000\rose_2\guidip\examples\1\Buseskod.isf opened.
Parsing EXM file...
Parsing OK.
File D:\Usr\dyplomanci2000\rose_2\guidip\examples\1\Buseskod.D1.rlf opened.
Parsing RLF file...
Parsing OK.
  
```

Buseskod.D1.rlf: Full View

NUM

Wymagania do sprawozdania

- Krótki opis problemu (raczej Twoje zrozumienie + cele postawione przez klienta)
- Informacje o danych
- Scenariusz metodyczny
- Wstępna ocena danych i ew. czyszczenie
- Analiza statystyk opisowych / korelacji itp./
- Ocena ważności atrybutów
 - Ranking
 - Ew. selekcja / redukcja
- Podsumowanie wiedzy klasyfikacyjnej
 - *Nie tylko* trafności
 - Próba interpretacji reprezentacji wiedzy
- Wnioski dla klienta

Źródło danych do problemu

- ❑ Dane zebrane przez dr hab. inż. J. Żak Wydziału maszyn Roboczych i Pojazdów Politechniki Poznańskiej
- ❑ J. Żak, J. Stefanowski. Determining maintenance activities of motor vehicles using rough sets approach. In *Proceedings of Euromaintenance'94 Conference*. Amsterdam 1994, 39 - 42.

O jakie autobusy chodziło

- ❑ Autosan H9-21 (prototyp 1969, produkowany lata 73-84) w Sanockiej Fabryce Autobusów .
- ❑ Podstawowy model autobusu międzymiastowego PKS w poprzednim okresie

