

Klasyfikowanie zagrożenia na drodze za pomocą regresji logistycznej

Marzena Nowakowska

Katedra Technologii Informatycznych

Wydział Zarządzania i Modelowania Komputerowego

Modelowanie logistyczne zagrożeń na drogach zamiejskich

Analiza wpływu cech opisujących sprawców wypadków na zagrożenie na drodze wyrażające się:

- **nieprawidłowym zachowaniem sprawcy**
- **rodzajem spowodowanego wypadku**
- statusem (ciężkością, dotkliwością) wypadku

Badania przeprowadzone na rozłącznych zbiorach danych o sprawcach:

- **kierujących pojazdami w wypadkach bez pieszych z jednym pojazdem: K_W1Pj**
- **kierujących pojazdami w wypadkach z co najmniej dwoma pojazdami: K_W2Pj**

Dane do analizy

Zbiory zawierają dane o sprawcach wypadków na drogach zamiejskich woj. świętokrzyskiego z lat 1999-2004

Z analizy wykluczono obserwacje zawierające dane brakujące w cechach uczestniczących w budowie modelu klasyfikacyjnego

Opis zbioru danych	Liczba obserwacji	
	uczestnicy	sprawcy
Kierujący w wypadkach bez pieszych z udziałem jednego pojazdu K_W1Pj	1834	1680
Kierujący w wypadkach bez pieszych z udziałem co najmniej dwóch pojazdów K_W2Pj	7472	3451

- Pjzd – rodzaj pojazdu
- PlKr – płeć kierującego
- AlkKr – nietrzeźwość kierującego
- GrWKr – grupa wiekowa kierującego
- ZchKr – zachowanie kierującego
- RdZd – rodzaj zdarzenia (wypadku)
- StsZd – status (dotkliwość) zdarzenia (wypadku)

Charakterystyka danych do analizy

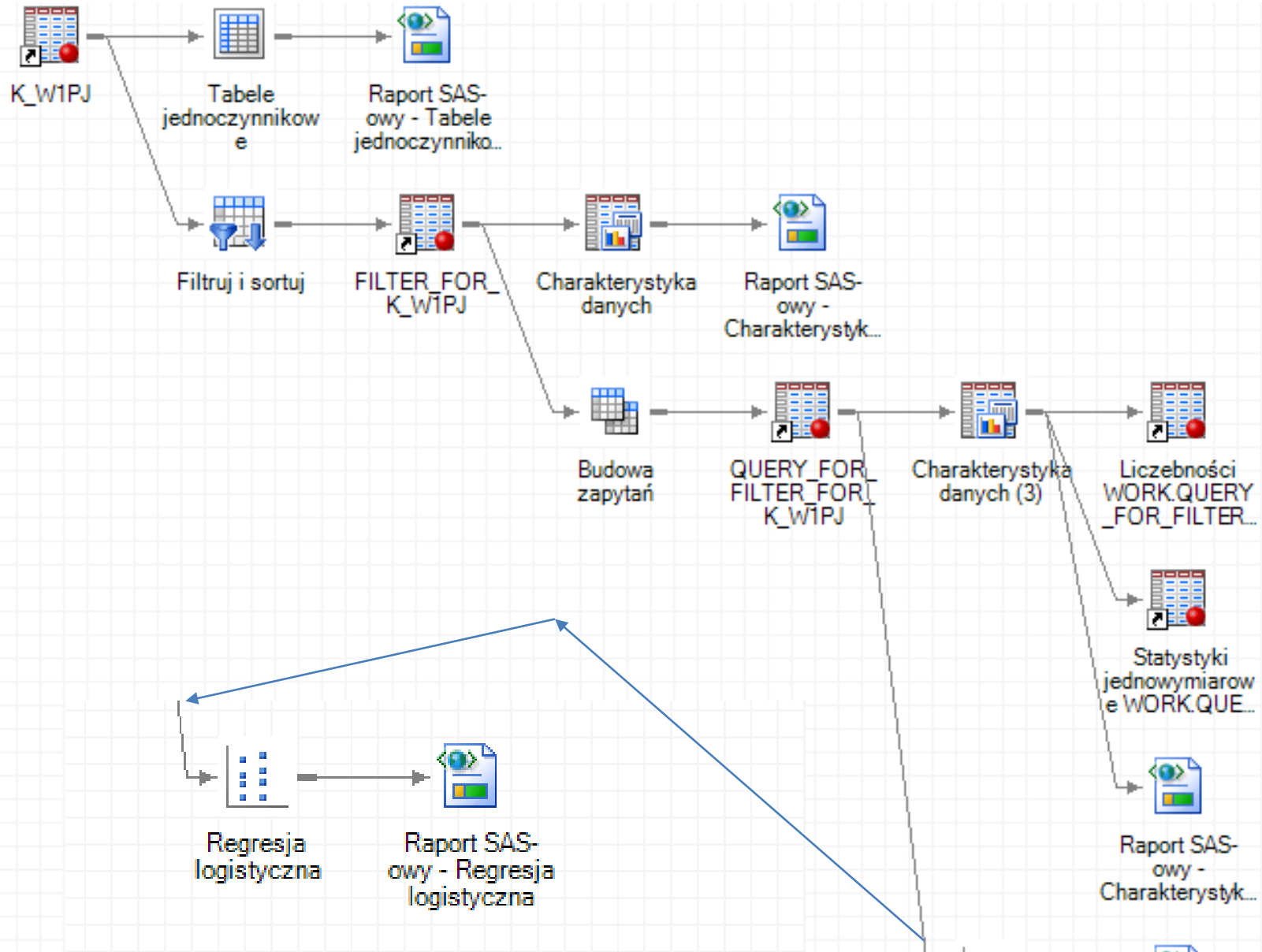
Cecha i jej dziedzina		K_W1Pj	K_W2j
Pjzd	R	4.5	15.7
	M	5.8	4.4
	O	75.2	60.4
	A	0.5	0.6
	C	10.5	15.4
	I	3.5	3.4
PIKr	K	12.7	12.1
	M	87.3	87.9
AlkKr	N	74.6	85.5
	T	25.4	14.5
GrWKr	1 - do 7	0	0.5
	2 - do 15	0.9	4.7
	3 - do 18	3.8	2.8
	4 - do 25	34.2	22.1
	5 - do 40	33.4	33.8
	6 - do 60	23.0	26.7
	7 - od 60	4.7	9.5

Kategoria odniesienia

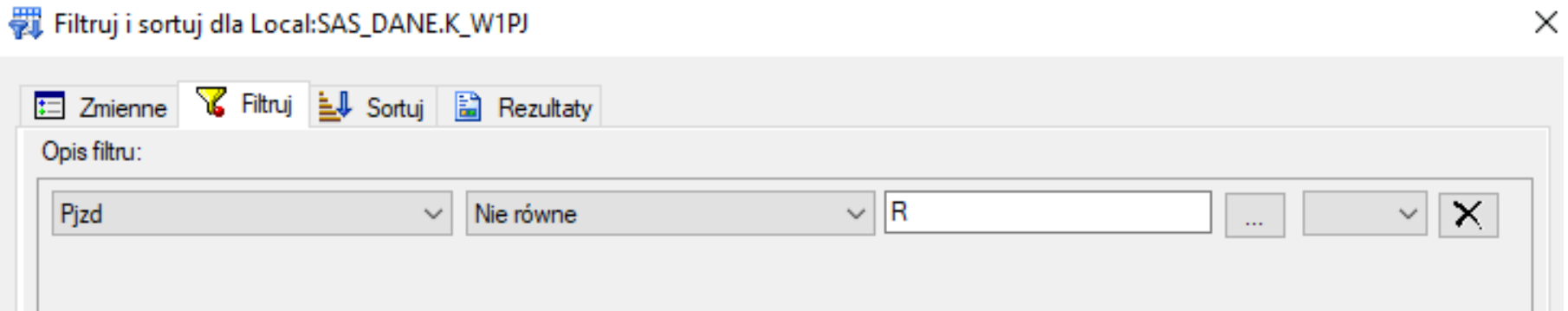
K_W1Pj	ZchKr	NdsPr	74.0
		Inn	17.0
		Pzst	9.0
	RzZd	NjDrzSpOb	44.8
K_W2Pj	ZchKr	WwPjzd	37.6
		Pzst	17.6
		NdzPrw	25.5
		NdsPr	23.6
		NprWO	15.5
		Inn	7.7
	RzZd	JzdNwS	6.8
		NzOd	5.5
		NprSkr	8.0
		Pzst	7.4
RzZd	ZdC	34.8	
	ZdB	31.8	
	ZdT	27.0	
	Pzst	6.4	

Kategoria sukcesu

Diagram przepływu informacji w SAS EG



Węzeł *Filtrowanie*



Odrzucenie pojazdów, które są rowerami

Uwaga: odrzucenie kategorii R powoduje, że ostatnią wartością jest O, definiując kategorię odniesienia.

Modyfikacja wartości zmiennych

Węzeł *Budowa zapytań*



Budowa zapytań dla Local:WORK.FILTER_FOR_K_W1PJ

Zapytanie: Budowa zapytań Wynik: WORK.QUERY_FOR_FILTER_FOR_K_W' Zmień...

Kolumny wyliczane Menedżer podpowiedzi Podgląd Narzędzia Opcje

Dodaj table Usuń

t1 (FILTER_FOR_K_W1PJ)

- Pjzd
- KPI
- KAlk
- GrWKr
- ZchKr
- Kolumny wyliczane
 - ZchK_kod

Wybierz dane Filtruj dane Sortuj dane

Nazwa kolumny	Kolumna ...	P..	Format	Szcze
Pjzd (Pjzd)	t1.Pjzd			
ZchK_kod	Wyliczane		\$5.	CASE
KPI (KPI)	t1.KPI			
KAlk (KAlk)	t1.KAlk			
GrWKr (GrWKr)	t1.GrWKr			
ZchKr (ZchKr)	t1.ZchKr			

Wybierz tylko wiersze bez powtórzeń

Uruchom Zapisz i zamknij Anuluj Pomoc

Modyfikacja wartości zmiennej – binaryzacja zmiennej Zch_Kr

Edytuj kolumnę wyliczaną



1 z 2 Podaj sposób zastępowania



Zastępowanie

Zastąp	na
= 'Inn'	'N_NdsPr'
= 'NdsPr'	'T_NdsPr'
= 'NprWO'	'N_NdsPr'
= 'NspKr'	'N_NdsPr'
= 'Rzn'	'N_NdsPr'

Inne wartości

Pozostałe wartości zastap:

Wartością bieżącą

Brakiem danych

Podaj wartość:

Ujmij wartość w cudzysłów

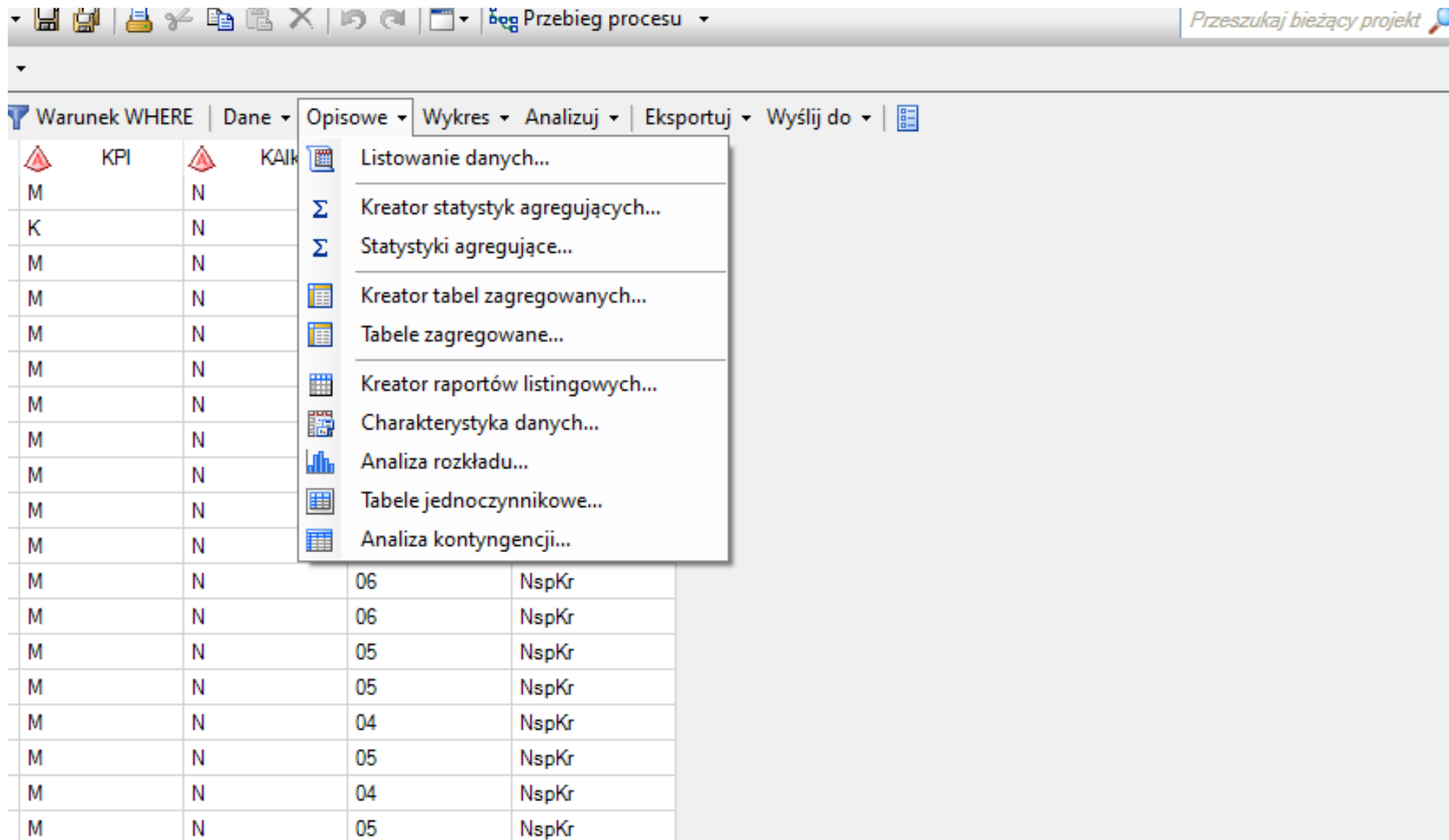
Typ kolumny

Znakowe

Numeryczne

< Wstecz | **Dalej >** | Koniec | Anuluj | Pomoc

Charakterystyka danych

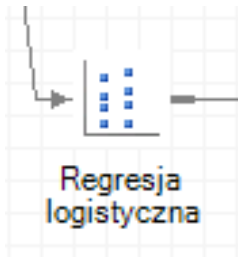


The screenshot shows a software interface with a menu bar at the top containing icons for file operations and a search bar with the text "Przeszukaj bieżący projekt". Below the menu bar, there is a toolbar with options: "Warunek WHERE", "Dane", "Opisowe", "Wykres", "Analizuj", "Eksportuj", and "Wyślij do". A context menu is open over the "Opisowe" option, listing several data analysis tools. The background shows a data table with columns "KPI" and "KAIK" containing values "M" and "N".

KPI	KAIK		
M	N		
K	N		
M	N		
M	N		
M	N		
M	N		
M	N		
M	N		
M	N		
M	N	06	NspKr
M	N	06	NspKr
M	N	05	NspKr
M	N	05	NspKr
M	N	04	NspKr
M	N	05	NspKr
M	N	04	NspKr
M	N	05	NspKr

- Listowanie danych...
- Kreator statystyk agregujących...
- Statystyki agregujące...
- Kreator tabel zagregowanych...
- Tabele zagregowane...
- Kreator raportów listingowych...
- Charakterystyka danych...
- Analiza rozkładu...
- Tabele jednoczynnikowe...
- Analiza kontyngencji...

Uruchamiane z poziomego podglądu pliku



Węzeł Regresja

Regresja logistyczna dla Local:WORK.QUERY_FOR_FILTER_FOR_K_W1PJ

czna dla Local:WORK.QUERY_FOR_FILTER_FOR

- Dane
- Model
- Odpowiedź
- Efekty
- Wybór
- Opcje
- Wykresy
- Prognozy
- Tytuły
- Właściwości

Dane

Źródło danych: Local:WORK.QUERY_FOR_FILTER_FOR_K_W1PJ
Filtr zadań: Brak

Zmienne przypisywane:

Nazwa
Pjzd
ZchK_kod
KPI
KAlk
GrWkr
ZchKr

Role zadania:

Zmienna zależna (Limit: 1)
Zch_kod
Zmienne ilościowe
Zmienne klasyfikujące
Pjzd
KPI
KAlk
GrWkr
Grupuj analizowane wg
Liczebność (Limit: 1)
Waga względna (Limit: 1)

Model > Odpowiedź

Typ odpowiedzi: Binarna

Typ modelu:
 logit
 probit
 clog-log
 glogit

Poziomy odpowiedzi dla Zch_kod:

N_Nds
T_Nds

Dopasuj model do poziomu: T_Nds

Wyniki – ocena modelu

Statystyki dopasowania		
Kryterium	Tylko wyraz wolny	Wyraz wolny i współzmienne
AIC	1739.936	1630.309
SC	1745.317	1694.880
-2 log L	1737.936	1606.309

R-kwadrat	0.0787	Maksymalnie przeskalowane R-kwadrat	0.1191
-----------	--------	-------------------------------------	--------

Testowanie globalnej hipotezy zerowej: BETA=0			
Testowanie	Chi-kwadrat	DF	Pr. > chi-kw.
Il. wiarygodn.	131.6266	11	<.0001
Mn. Lagrange'a	153.9262	11	<.0001
Walda	96.8598	11	<.0001

Analiza efektów typu 3			
Efekt	DF	Chi-kwadrat Walda	Pr. > chi-kw.
Pjzd	4	71.2317	<.0001
KPI	1	0.1803	0.6712
KAlk	1	0.3175	0.5731
GrWKR	5	19.6144	0.0015

Wyniki – analiza efektów typu 3

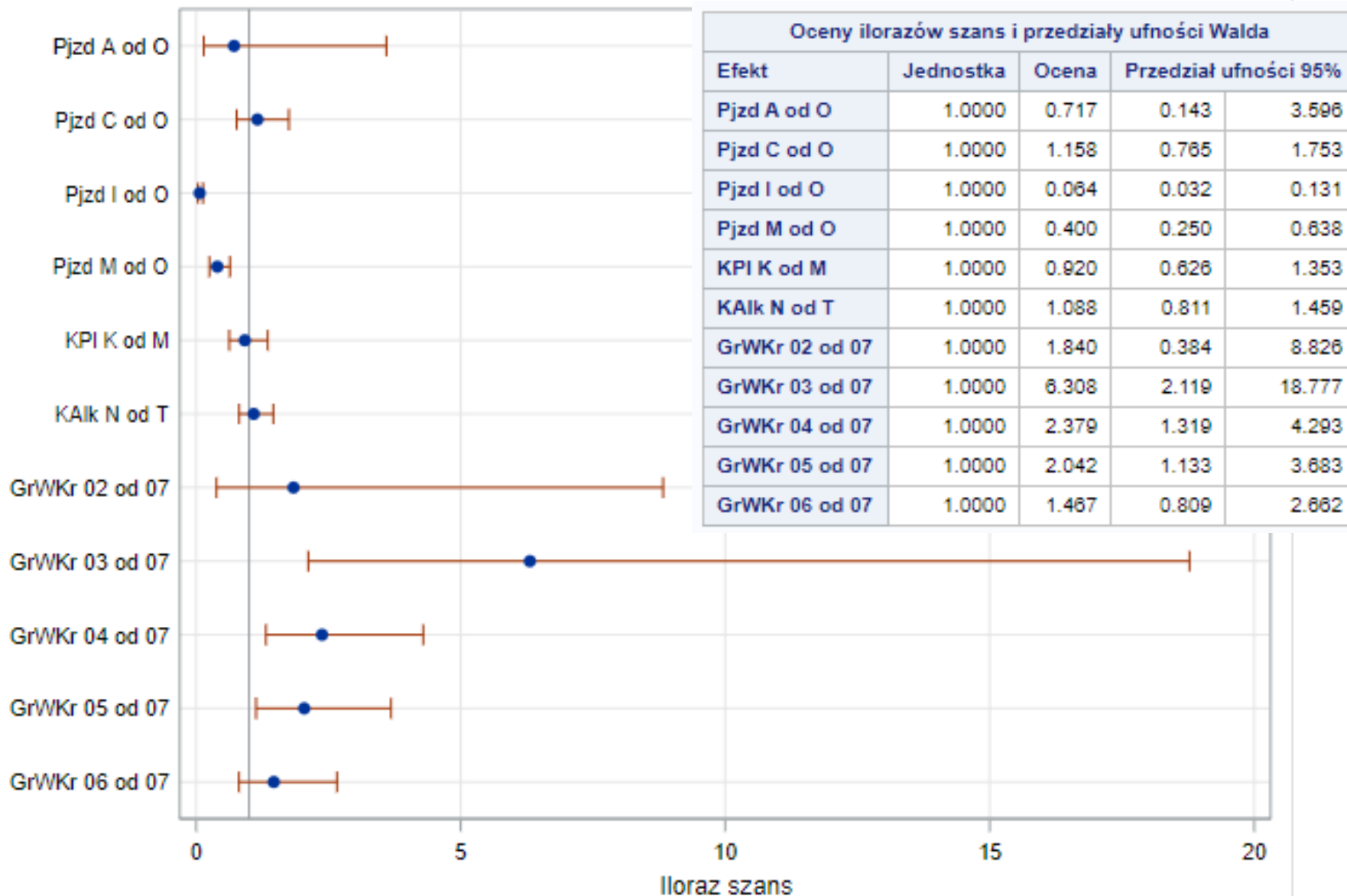
Analiza efektów typu 3			
Efekt	DF	Chi-kwadrat Walda	Pr. > chi-kw.
Pjzd	4	71.2317	<.0001
KPI	1	0.1803	0.6712
KAlk	1	0.3175	0.5731
GrWkR	5	19.6144	0.0015

Analiza ocen maksymalnej wiarygodności						
Parametr		DF	Ocena	Błąd standardowy	Chi-kwadrat Walda	Pr. > chi-kw.
Intercept		1	0.6125	0.2598	5.5568	0.0184
Pjzd	A	1	0.4360	0.6624	0.4334	0.5103
Pjzd	C	1	0.9162	0.2438	14.1204	0.0002
Pjzd	I	1	-1.9740	0.3324	35.2736	<.0001
Pjzd	M	1	-0.1477	0.2587	0.3259	0.5681
KPI	K	1	-0.0418	0.0984	0.1803	0.6712
KAlk	N	1	0.0422	0.0748	0.3175	0.5731
GrWkR	02	1	-0.1260	0.6331	0.0396	0.8422
GrWkR	03	1	1.1058	0.4211	6.8951	0.0086
GrWkR	04	1	0.1308	0.1836	0.5077	0.4761
GrWkR	05	1	-0.0219	0.1861	0.0138	0.9084
GrWkR	06	1	-0.3528	0.1922	3.3702	0.0664

**Wyniki –
ocena
parametrów**

Wyniki – ilorazy szans

Ilorazy szans z przedziałem ufności Walda 95%



Wyniki – interpretacja ilorazów szans

Wymodelowane prawdopodobieństwo wynosi $ZchK_kod='T_Nds'$.

Oceny ilorazów szans i przedziały ufności Walda				
Efekt	Jednostka	Ocena	Przedział ufności 95%	
Pjzd A od O	1.0000	0.717	0.143	3.596
Pjzd C od O	1.0000	1.158	0.765	1.753
Pjzd I od O	1.0000	0.064	0.032	0.131
Pjzd M od O	1.0000	0.400	0.250	0.638
KPI K od M	1.0000	0.920	0.626	1.353
KAlk N od T	1.0000	1.088	0.811	1.459
GrWKR 02 od 07	1.0000	1.840	0.384	8.826
GrWKR 03 od 07	1.0000	6.308	2.119	18.777
GrWKR 04 od 07	1.0000	2.379	1.319	4.293
GrWKR 05 od 07	1.0000	2.042	1.133	3.683
GrWKR 06 od 07	1.0000	1.467	0.809	2.662

Szansa, że zachowaniem kierującego będącym przyczyną wypadku jest niedostosowanie prędkości do warunków ruchu jest:

- dla motocykla o 60% mniejsze niż dla samochodu osobowego ($IS=0,40$)
- dla osób z 3-ciej grupy wiekowej (15-18 lat) ponad 6 razy większe niż dla osób z 7-ej grupy wiekowej (wiek 60+); grupa 60+ to najbezpieczniejsza grupa kierujących o rozważanym zachowaniu

Jakość klasyfikacji

Tabela klasyfikacji									
Poziom prawd.	Poprawnie		Niepoprawnie		Procenty				
	Wystąpienie	Nie- wyst.	Wystąpienie	Nie- wyst.	Poprawnie	Czu- łość	Specy- ficzność	Progn. dod.	Progn. ujemn.
0.500	1223	53	319	10	79.5	99.2	14.2	79.3	84.1

Jeżeli prawdopodobieństwo sukcesu jest co najmniej 0,5 → obserwacja jest sukcesem

Wystąpienie oznacza sukces ($TPR=1223/1233=99,2\%$), niewystąpienie – porażkę ($TNR=53/327=14,2\%$)

Progn. dod oznacza precyzję przewidywania pozytywnego: $PPV=1223/1542=79,3\%$

Progn. ujemn oznacza precyzję przewidywania negatywnego: $NPV=53/63=84,1\%$

	Prognozowane porażki (ujemne)	Prognozowane sukcesy (dodatnie)	Liczebności obserwowane
Obserwowane porażki (negatywne)	53	319	372
Obserwowane sukcesy (pozytywne)	10	1223	1233

Niwelowanie nierównomierności rozkładu zmiennej celu

Dla większości algorytmów klasyfikacji duże różnice w częstościach kategorii sukcesu i porażki nie wpływają dobrze na jakość uzyskiwanych wyników. W skrajnych przypadkach, gdy przewaga jakiejś kategorii jest bardzo duża, modele prognostyczne mogą wszystkie obserwacje zaliczać do tej właśnie wartości, dostarczając formalnie dobrą ocenę modelu, w którym niewielki błąd klasyfikacji jest równy odsetkowi skrajnie nielicznych kategorii.

Remedium - próbkowanie poprzez losowanie warstwowe (wg kategorii zmiennej objaśnianej) na dwa sposoby:

- pobrać pełną reprezentację z kategorii mniej licznej oraz tyle obserwacje z warstwy kategorii częstszej, aby dostosować jej liczebność do liczby obserwacji wartości najrzadszej,
- pobrać określoną liczbę obserwacji o kategorii liczniejszej oraz powielić rekordy z warstwy rzadszej dostosowując jej częstość do ww. liczebności

