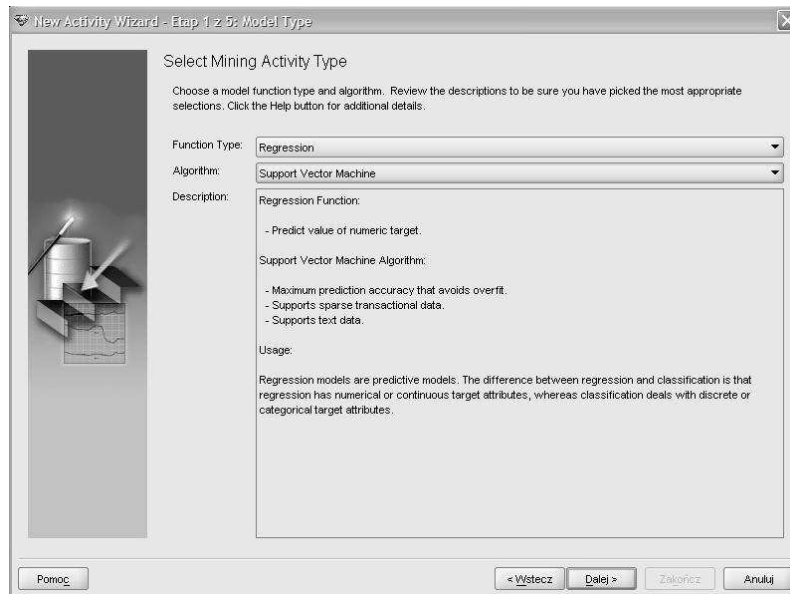


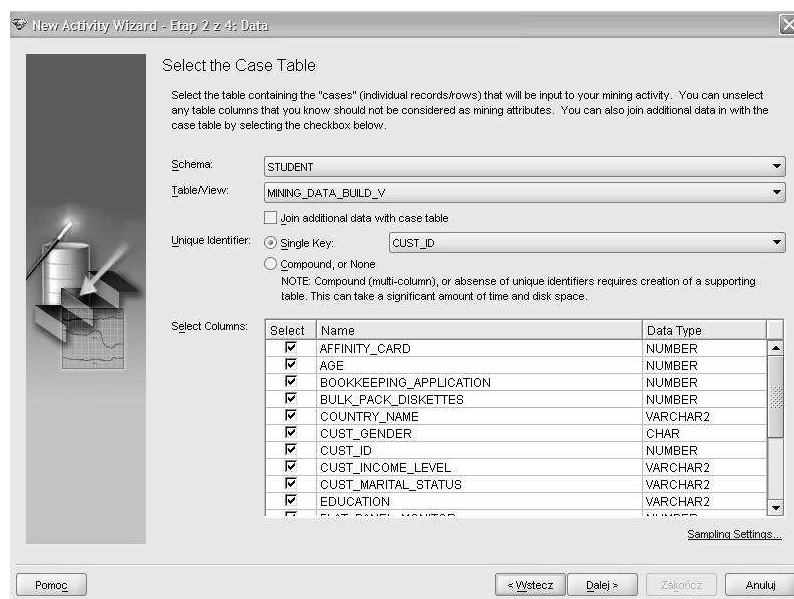
Laboratorium 11

Regresja SVM.

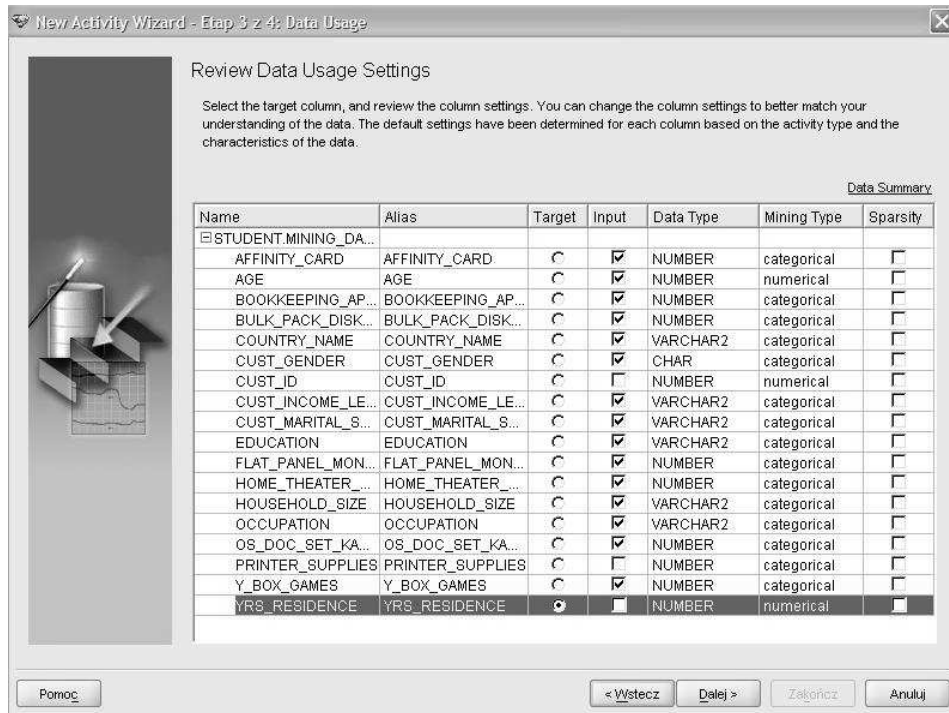
1. Uruchom narzędzie Oracle Data Miner i połącz się z serwerem bazy danych.
2. Z menu głównego wybierz Activity→Build. Na ekranie powitalnym kliknij przycisk Dalej>.
3. Z listy Function Type wybierz Regression. Rozwiń listę Algorithm i wybierz z niej algorytm Support Vector Machines. Kliknij przycisk Dalej>.



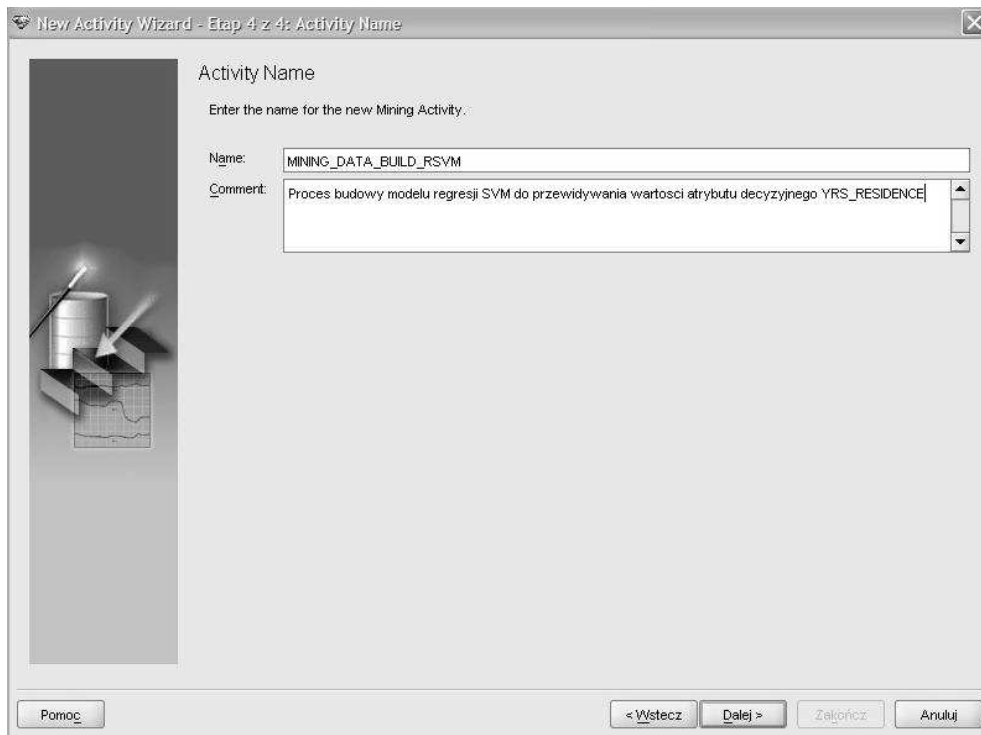
4. Wskaż schemat STUDENT i tabelę MINING_DATA_BUILD_V jako źródło danych do eksploracji. Jako klucz podstawowy wskaż atrybut CUST_ID. Kliknij przycisk Dalej>.



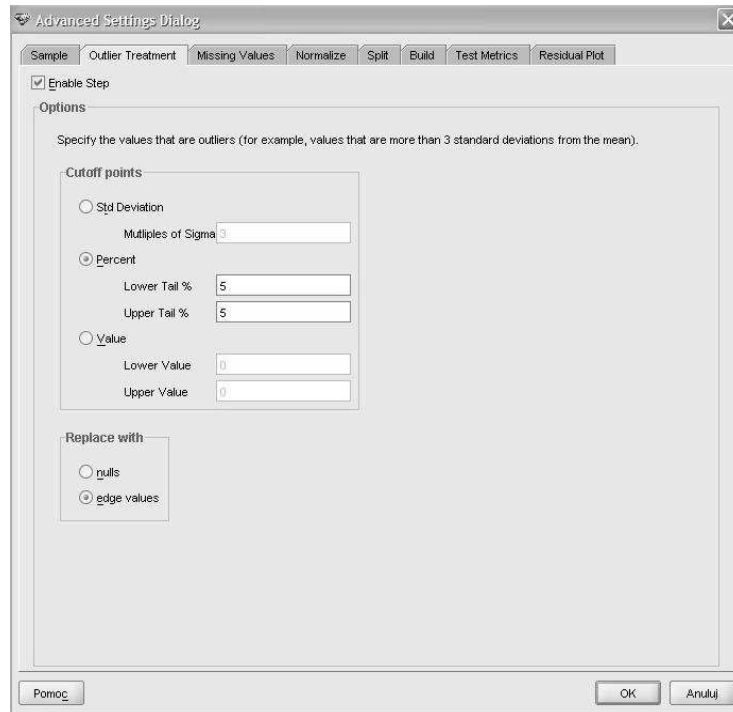
- Jako atrybut decyzyjny zaznacz atrybut YRS_RESIDENCE (pole radiowe w kolumnie Target). Zwróć uwagę, aby wartość atrybutu decyzyjnego została wyłączona z budowy klasyfikatora (pole wyboru Input dla atrybutu YRS_RESIDENCE musi być odznaczone). Upewnij się, że atrybuty CUST_ID i PRINTER_SUPPLIES są wyłączone z eksploracji (są bezwartościowe i nie niosą żadnej informacji). Kliknij przycisk **Dalej>**.



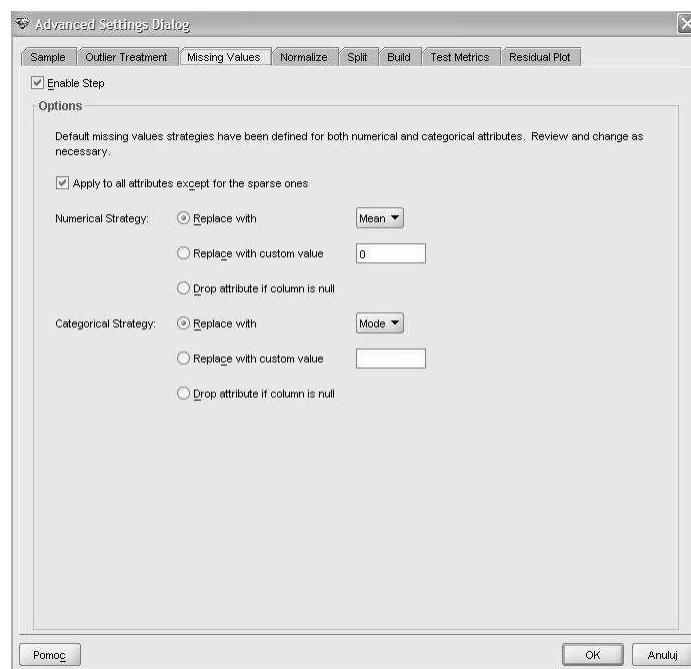
- Podaj nazwę i krótki opis procesu eksploracji. Kliknij przycisk **Dalej>**.



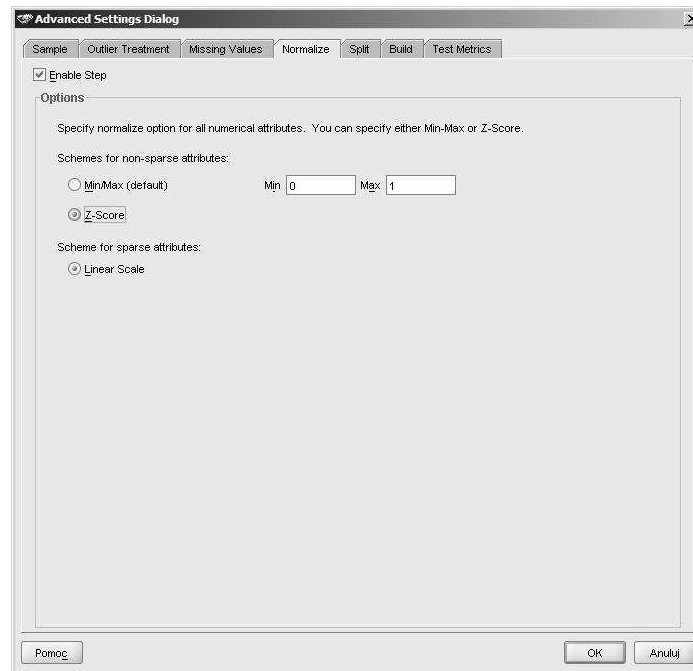
7. Kliknij przycisk **Advanced Settings**. Upewnij się, że na zakładce **Sample** opcja próbkowania jest wyłączona (pole wyboru **Enable Step** jest odznaczone). Przejdź na zakładkę **Outlier Treatment**. Algorytm SVM jest bardzo czuły na występowanie osobliwości. Oznacz jako osobliwości po 5% wartości z każdego końca przedziału wartości zastępując usuwane osobliwości wartościami brzegowymi.



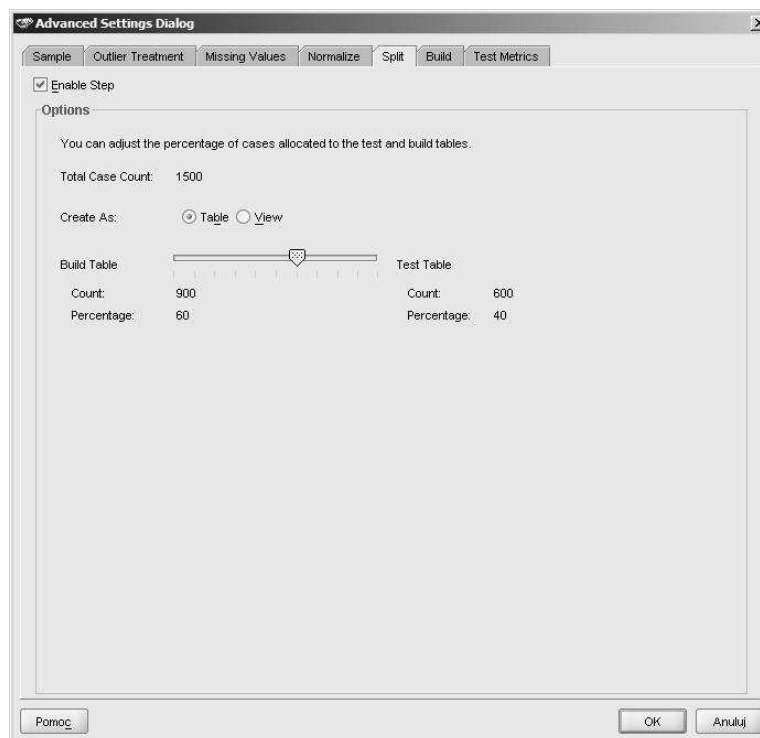
8. Przejdź na zakładkę **Missing Values**. Upewnij się, że przetwarzanie brakujących wartości jest włączone (pole wyboru **Enable Step** musi być zaznaczone). Wartości puste występujące w atrybutach numerycznych zamień na wartość średnią (**Mean**), a wartości puste występujące w atrybutach kategoriycznych zamień na wartość modalną (**Mode**).



9. Przejdź na zakładkę Normalize. Algorytm SVM wymaga, aby wszystkie atrybuty numeryczne były znormalizowane. Jako metodę normalizacji wybierz wyrażenie wartości w liczbie odchyłeń standardowych od średniej (zaznacz pole radiowe Z-Score).

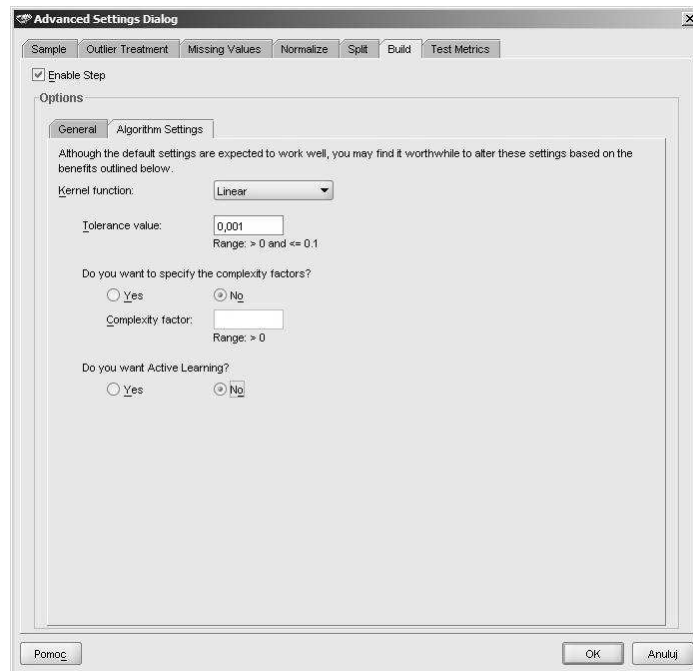


10. Przejdź na zakładkę Split. Dokonaj podziału zbioru wejściowego na zbiór uczący i testujący w proporcjach 60%-40%, podział powinien wykorzystywać tabelę.

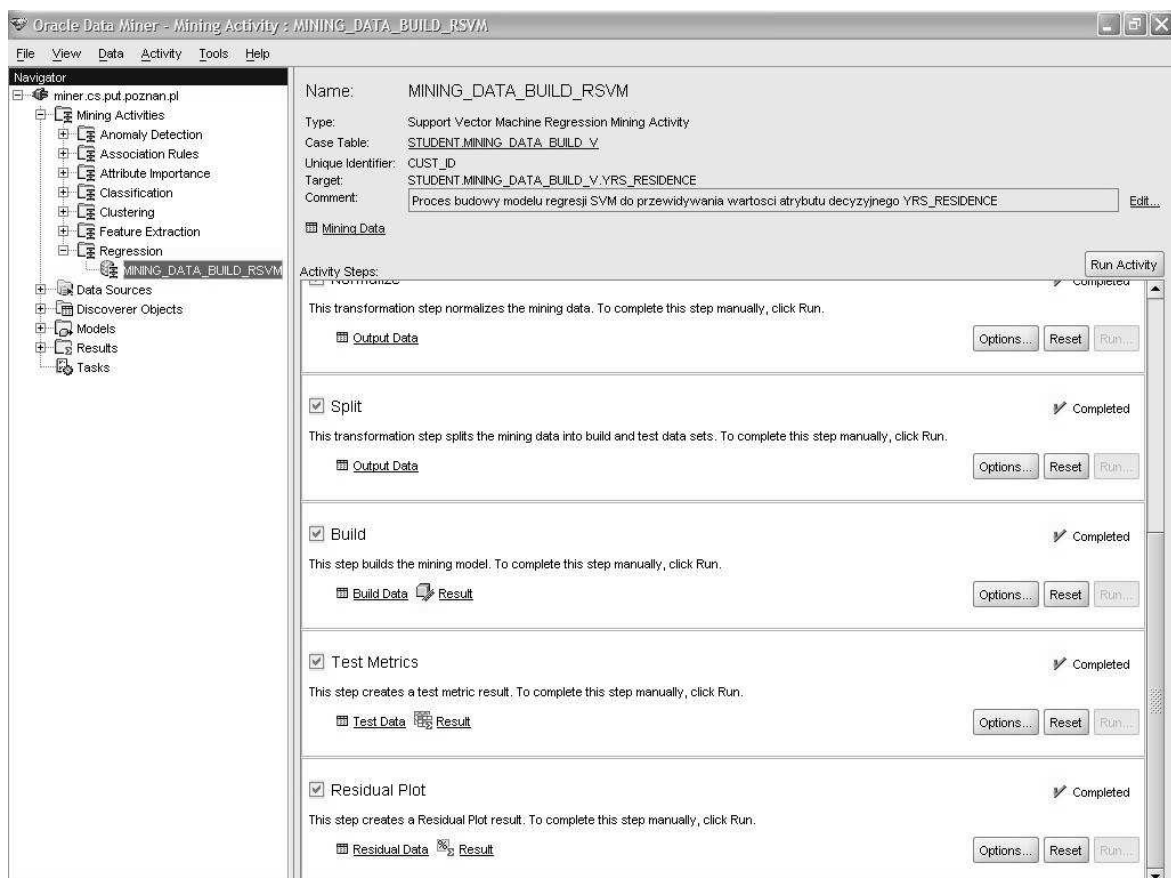


11. Przejdź na zakładkę Build. Upewnij się, że algorytm będzie się starał osiągnąć maksymalną średnią dokładność (w polu Accuracy Goal wybierz opcję Maximum Average Accuracy). Kliknij na zakładkę Algorithm Settings. Jako rodzaj funkcji

jądrowej wskaź funkcję liniową. Koniecznie wyłącz opcję aktywnego uczenia (pole radiowe Do you want Active Learning?, opcja No).



12. Kliknij przycisk **OK**. Upewnij się, że opcja Run upon finish jest włączona. Kliknij przycisk **Zakończ**.



13. Kliknij na odnośnik Result w bloku Build. Współczynniki przy każdej wartości predyktorów definiują hiperpłaszczyznę najlepiej separującą instancje należące do klas decyzyjnych. Zauważ, że uzyskany wynik w praktyce nie poddaje się naturalnej interpretacji i stanowi rodzaj „czarnej skrzynki”.

Result Viewer: MINING_DATA_B60198_SV

File Help

Coefficients Results Build Settings Task

Bias: 1,78120114

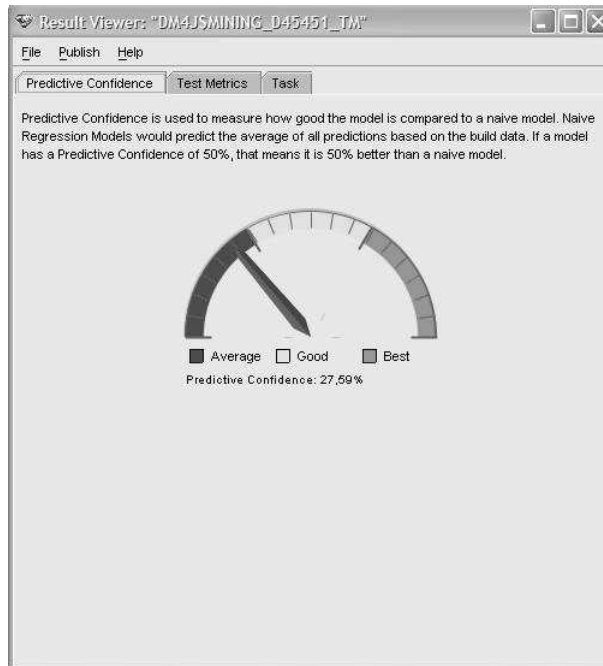
Coefficients

Fetch Size: 100 Refresh Unscale Filter Zamknij

Attribute Name	Value	Coefficient
COUNTRY_NAME	Denmark	1,8808374906
EDUCATION	PhD	0,8775112834
EDUCATION	Profsc	0,7861432271
CUST_MARITAL_STATUS	Married	0,7416823567
EDUCATION	12th	0,5930864552
EDUCATION	5th-6th	0,4253914143
EDUCATION	9th	0,4184288846
CUST_MARITAL_STATUS	Separ.	0,3747753385
CUST_MARITAL_STATUS	Divorc.	0,3675041938
BOOKKEEPING_APPLICATION	1	0,3501127495
HOUSEHOLD_SIZE	6-8	0,3387982934
COUNTRY_NAME	Japan	0,3235127788
CUST_INCOME_LEVEL	A: Below 30,000	0,3039808498
OCCUPATION	Exec.	0,2847642904
COUNTRY_NAME	Poland	0,2823985171
HOUSEHOLD_SIZE	2	0,2636129617
OCCUPATION	Machine	0,2410771706
CUST_MARITAL_STATUS	NeverM	0,2279862283
OCCUPATION	Sales	0,2030708858
HOME_THEATER_PACKAGE	1	0,1954888611
OCCUPATION	Transp.	0,1797580871
COUNTRY_NAME	Canada	0,1771043091
CUST_INCOME_LEVEL	H: 150,000 - 169,999	0,1718609108
CUST_MARITAL_STATUS	Mabsent	0,1441433606
EDUCATION	7th-8th	0,1394896892
HOUSEHOLD_SIZE	9+	0,1375129912
AFFINITY_CARD	1	0,1199630413
CUST_INCOME_LEVEL	E: 90,000 - 109,999	0,1056427656
CUST_INCOME_LEVEL	I: 170,000 - 189,999	0,1050348026
OCCUPATION	Cleric.	0,1000856607
OCCUPATION	Crafts	0,0997547056
OCCUPATION	Farming	0,0957994000
CUST_GENDER	M	0,0837532800

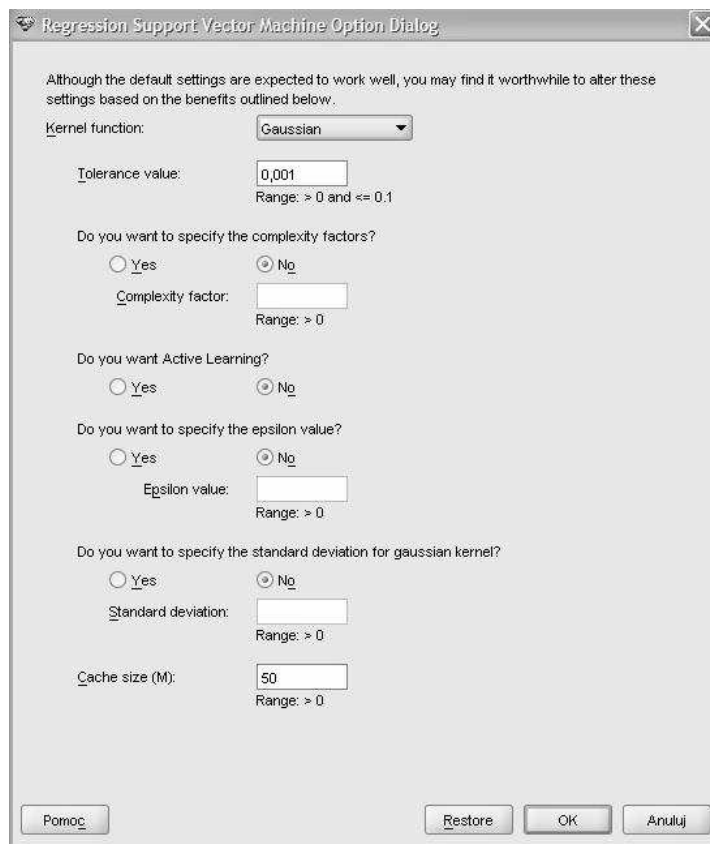
Sort coefficients based on absolute values

14. Zamknij okno z wynikami budowy klasyfikatora i powróć do głównego okna. Kliknij odnośnik Result w bloku Test Metrics. Na zakładce Predictive Confidence przedstawiona jest dokładność klasyfikatora liczona względem naiwnego klasyfikatora 0-R, który zawsze przewiduje najczęstszą wartość atrybutu decyzyjnego.



15. Powróć do głównego okna programu. Zaobserwuj zmianę jakości wygenerowanego klasyfikatora po korekcie parametrów algorytmu. Kliknij przycisk **Reset** w bloku Build (spowoduje to zresetowanie tego i wszystkich kolejnych kroków procesu odkrywania wiedzy).

16. Kliknij przycisk Options w bloku Build. Przejdź na zakładkę Algorithm Settings. Zmień rodzaj funkcji jądrowej na Gaussowską. Upewnij się, że opcja aktywnego uczenia się jest wyłączona.



17. Powróć do głównego okna programu. Kliknij przycisk Run Activity (prawy górny okna). Po zakończeniu się procesu odkrywania wiedzy kliknij odnośnik Result w bloku Test Metrics. Czy nowy klasyfikator jest lepszy czy gorszy od poprzedniego?

Ćwiczenie samodzielne

Wykorzystaj model stworzony za pomocą skryptu svm.reg.plsql do predykcji wieku klientów z Włoch. Jako źródło danych wykorzystaj tabelę MINING_DATA_APPLY_V. Pamiętaj o poddaniu danych źródłowych tym samym transformacjom, jakim były poddane dane, na podstawie których zbudowano model. Wykorzystując narzędzie GnuPlot dokonaj wizualizacji otrzymanych wyników.

Postaraj się uzyskać rezultat podobny do zamieszczonego poniżej:

